Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A survey on semantic processing techniques

Rui Mao [a], Kai He [c], Xulang Zhang [b], Guanyi Chen [d,e], Jinjie Ni [b], Zonglin Yang [a], Erik Cambria [a,*]

[a] Continental-NTU Corporate Lab, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[b] School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[c] Saw Swee Hock School of Public Health, National University of Singapore, 117549, Singapore
[d] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, 382 Xiongchu Avenue, 430079, Wuhan, China
[e] School of Computer Science, Central China Normal University, 382 Xiongchu Avenue, 430079, Wuhan, China

ARTICLE INFO

Keywords:
Semantic processing
Word sense disambiguation
Anaphora resolution
Named entity recognition
Concept extraction
Subjectivity detection

ABSTRACT

Semantic processing is a fundamental research domain in computational linguistics. In the era of powerful pre-trained language models and large language models, the advancement of research in this domain appears to be decelerating. However, the study of semantics is multi-dimensional in linguistics. The research depth and breadth of computational semantic processing can be largely improved with new technologies. In this survey, we analyzed five semantic processing tasks, e.g., word sense disambiguation, anaphora resolution, named entity recognition, concept extraction, and subjectivity detection. We study relevant theoretical research in these fields, advanced methods, and downstream applications. We connect the surveyed tasks with downstream applications because this may inspire future scholars to fuse these low-level semantic processing tasks with high-level natural language processing tasks. The review of theoretical research may also inspire new tasks and technologies in the semantic processing domain. Finally, we compare the different semantic processing techniques and summarize their technical trends, application trends, and future directions.

1. Introduction

Semantics is a linguistic term, generally referring to the meaning of language. Unlike syntax which studies the structure of sentences [1], the significance of semantics lies in its ability to aid our comprehension of how meaning is conveyed through words, phrases, and sentences, as well as how language is used to express various ideas, thoughts, and emotions. Language is one of the important carriers of meanings. However, the term “meaning” encompasses multiple aspects of language.

Palmer [2] argued that there is a lack of consensus regarding the nature of “meaning”, e.g., which components should be considered part of semantics, and how it should be characterized. Thus, the study of “semantics” is also multi-dimensional in academia. The evolution of semantic research reflects the rich connotation of semantics in linguistics. At the early stage, much attention is given to the study of lexical semantics. The first English dictionary, *Robert Cawdrey’s Table Alphabeticall*, dates back to 1604 [3]. The construction of dictionaries, e.g., *The Oxford English Dictionary* [4] became one of the most significant symbols of lexical semantic research achievements.

Research on lexical semantics covers word senses, polysemy, word formation, contrastive lexical semantics, and more. Next, another important research dimension of semantics emerged, termed structural semantics. Structural semantics emphasizes the analysis of sentence structures, including the relationships between words and the ways in which words contribute to the meaning of a sentence. The study of structural semantics includes but is not limited to analyzing the meaning of words by syntax, grammar, and pragmatics. Structural semantics elevates the study of semantics from the word level to the sentence level. The later cognitive semantics further enrich the connotation of semantics.

The tenets of cognitive semantics posit that the faculty of language is intricately intertwined with the broader cognitive capacity of human beings [5]. In other words, semantics is a reflection of how humans understand and make sense of the world around them. Under cognitive semantics, researchers extend to frame semantics (semantics is the reflection of encyclopedic knowledge), situation semantics (semantics reflects the relationships between situations) [6], conceptual semantics (semantics reflects the structural perception of concepts) [7], and more. Fig. 1 summarizes partial semantic research domains in linguistics.

[*] Corresponding author.
E-mail addresses: rui.mao@ntu.edu.sg (R. Mao), kai_he@nus.edu.sg (K. He), xulang001@e.ntu.edu.sg (X. Zhang), g.chen@ccnu.edu.cn (G. Chen), jinjie001@e.ntu.edu.sg (J. Ni), zonglin001@e.ntu.edu.sg (Z. Yang), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.inffus.2023.101988
Received 11 May 2023; Received in revised form 24 August 2023; Accepted 25 August 2023
Available online 1 September 2023

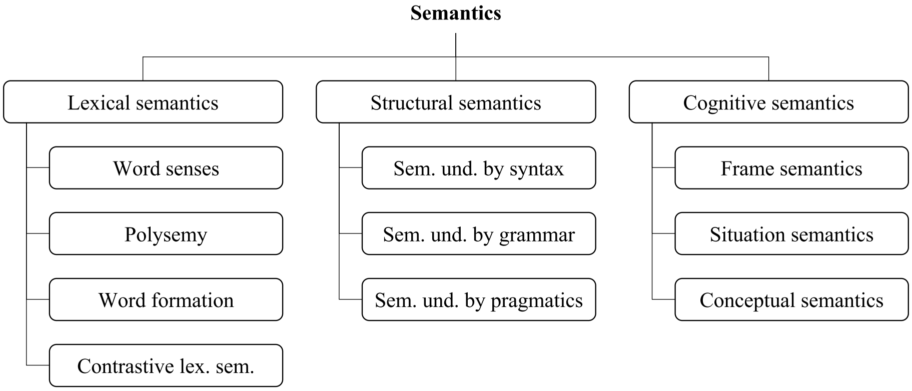


Fig. 1. Semantic research domains in linguistics. Lex. denotes lexical; sem. denotes semantics; und. denotes understanding.

Table 1
The surveyed semantic processing tasks and their downstream applications. F denotes that the technique yielded features for a downstream task model; P denotes that the technique was used as a parser; E denotes that the technique improved the explainability for a downstream task. WSD denotes word sense disambiguation. AR denotes anaphora resolution. NER denotes named entity recognition. CE denotes concept extraction. SD denotes subjectivity detection.

Downstream tasks	WSD	AR	NER	CE	SD
Sentiment computing	F, P, E	F		F, P, E	P
Information retrieval	E			F, E	P
Machine translation	F, P, E	F, E			
Summarization		F			
Textual entailment		F			
Knowledge graph construction			P		
Recommendation systems			F, P, E		
Dialogue systems			P, E	F, P	
Commonsense explanation generation				F, E	
Hate speech detection					F, P
Question & answering systems					F, P

The development of automatic semantic processing techniques has largely facilitated semantic research. Many useful tools and knowledge bases[1]were developed for word sense disambiguation, anaphora resolution, named entity recognition, concept extraction, and subjectivity detection. These tools are the embodiment of many theoretical ideas in semantics. For example, word sense disambiguation is an important task in lexical semantics. Anaphora resolution elucidates the relationship between the anaphor, which is the repetition of a reference, and its antecedent, which is the earlier mention of the entity. Anaphora resolution determines the structural semantics of the anaphor. Named entity recognition categorized named entities in texts by conceptually related classes, e.g., names, and locations. Similarly, concept extraction and subjectivity detection tasks also embody the cognitive properties of semantics.

In addition to improving semantic research, semantic processing techniques can also help other downstream natural language processing (NLP) tasks with more complexity (see Table 1). For example, subjectivity detection can be an upstream task of sentiment analysis, because subjective expressions can be further categorized by positive, negative, and neutral expressions with different opinionated intensities. The semantic processing techniques that have been reviewed possess a range of potential applications, including the ability to generate features that are effective, as well as to be used as a parser in order to obtain desired categories of text. Additionally, these techniques have the potential to improve the explainability of downstream applications.

The emergence of pre-trained language models (PLMs) has greatly enhanced the semantic representation capabilities of deep learning models and the ability to fit downstream tasks [8–10]. Some large language models (LLMs), e.g., GPT-4[2] and Bard[3] even realize the functions of multiple complex NLP tasks by the means of dialogue, such as question answering, translation, and text summarization. Many semantic processing studies have gradually faded out of the field of NLP. Then, in the era of PLMs and LLMs, an intuitive question is what is the motivation for studying semantic processing techniques?

As mentioned before, semantics reflects the multiple aspects of language. Besides understanding word senses, semantics is also the entrance to understanding the mechanism, and perception of language. Language intelligence encompasses more than just achieving a level of accuracy that is equivalent to or surpasses human accuracy for specific tasks. It also entails the capacity to unveil the nature of language and investigate the cognitive processes that underlie language. Much aforementioned semantic research in the context of linguistics has not been explored in computational linguistics to our best knowledge. Thus, we are motivated to propose a survey on semantic processing techniques to encourage future scholars that can expand the depth and breadth of semantic research, leading the public attention from the application value of NLP techniques to the research value of computational linguistics. Nevertheless, we also highlight the fusion of low-level semantic processing techniques and high-level NLP techniques to demonstrate the application value of semantic processing techniques in different domains.

Given the broadness of semantics, our survey scope lies in semantic processing techniques for word sense disambiguation, anaphora resolution, concept extraction, named entity recognition, and subjectivity detection. This is because these low-level semantic processing tasks

[1] A knowledge base normally refers to a collection of organized information that is machine-readable, and supportive for an intelligent system.

[2] https://openai.com/product/gpt-4

[3] https://bard.google.com/

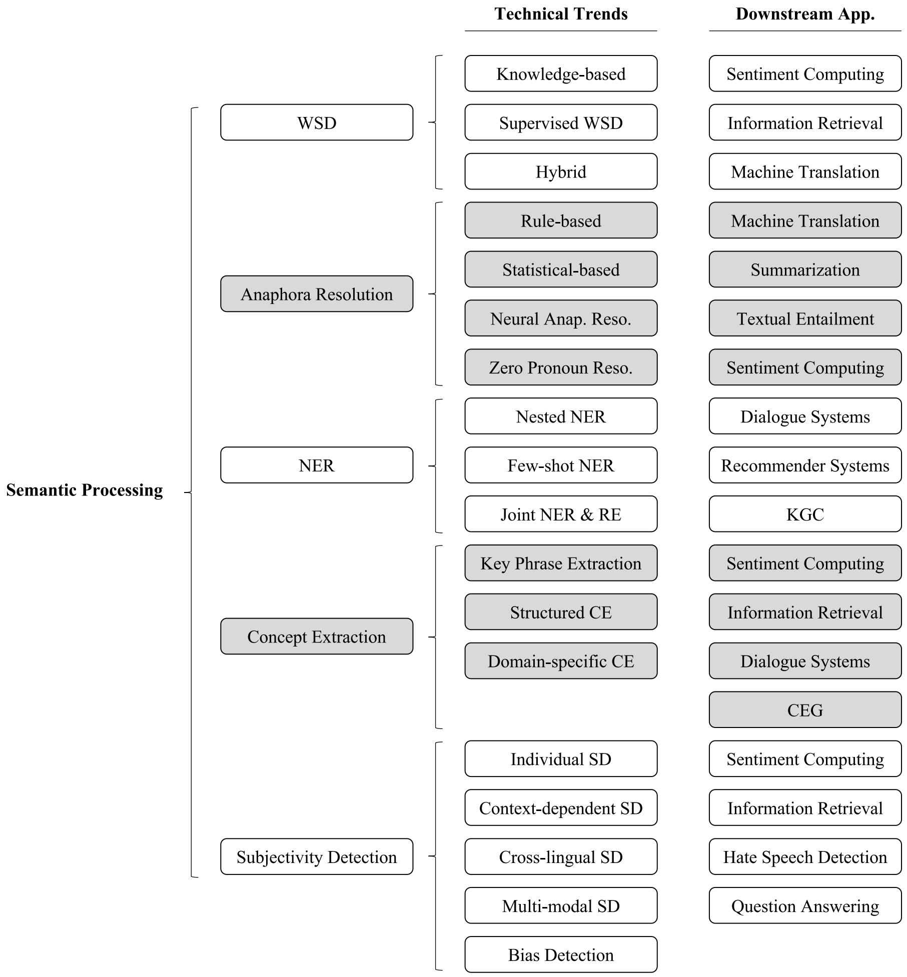


Fig. 2. The summary of technical trends and downstream applications of surveyed semantic processing tasks. KGC denotes knowledge graph construction. CEG denotes commonsense explanation generation. RE denotes relation extraction.

reflect different aspects of semantics. In addition, there were many research works on these tasks in the field of computational linguistics. We focus on low-level semantic processing tasks, rather than high-level semantic processing tasks, e.g., sentiment analysis and natural language inference, because they provide fundamental building blocks for both high-level semantic processing tasks and higher-level NLP tasks.

Multiple semantic processing techniques were rarely surveyed in the same article. Salloum et al. [11] surveyed several high-level semantic processing tasks, e.g., latent semantic analysis, explicit semantic analysis, and sentiment analysis. Compare to the work of Salloum et al. [11], our survey includes the latest research in low-level semantic processing techniques. Compare to the latest semantic processing surveys focusing on specific tasks [12–16], we additionally reviewed important theoretical research and downstream task applications in these domains. These contents can help readers better understand the foundation of semantic research in linguistics, as well as potential application scenarios. More importantly, theoretical research shows the big picture of a semantic processing task, which may inspire different research tasks in the computational linguistic community. The collection of multiple semantic processing techniques is helpful for readers to have a comprehensive understanding of a large field, inspiring more fusion research across different domains. Theoretical research of other tasks has the potential to inspire fresh perspectives among researchers who have been concentrating on a specific semantic research task.

The contribution of this survey is threefold:

- We survey recent semantic processing techniques, annotation tools, datasets, and knowledge bases for five low-level semantic processing tasks.
- We highlight important theoretical research, and downstream applications to encourage deeper and wider research in the semantic processing domain upon the currently established task setups.
- We compare different semantic processing techniques, delineate their technical and application trends, and put forth potential avenues for future research in this domain.

In the following sections, we introduce different semantic processing techniques, e.g., word sense disambiguation (Section 2), anaphora resolution (Section 3), named entity recognition (Section 4), concept extraction (Section 5), and subjectivity detection (Section 6). We discuss the interactions between the surveyed tasks and the impacts of deep learning and LLMs on semantic processing in Section 7. Finally, we conclude this survey in Section 8. Each task is structured by theoretical research, annotation schemes, datasets, knowledge bases, evaluation metrics, methods, downstream applications, and a summary. Fig. 2 demonstrates the taxonomy of methods and downstream applications of each task in this survey.

2. Word sense disambiguation

The complexity of human language is difficult for machines to understand it. One of the challenges is the ambiguity of word senses. In natural language, a word may have multiple senses, given different contexts. Consider the following example:

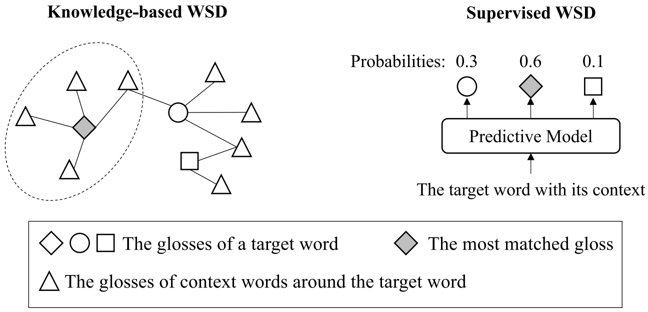

Knowledge-based WSD
Supervised WSD
Probabilities: 0.3 0.6 0.1
Predictive Model
The target word with its context
◇ ○ □ The glosses of a target word
◆ The most matched gloss
△ The glosses of context words around the target word

Fig. 3. Simplified examples of the knowledge-based and supervised WSD.

(1) He got his shoes wet as he walked along the **bank**.

According to the Oxford English Dictionary, the major senses of "bank" include (a) *an organization that provides various financial services, for example keeping or lending money*; (b) *the side of a river, canal, etc. and the land near it*. With the context, humans can easily know that "bank" here refers to the sense (b). However, it is challenging for machines to do so because the interpretation made by humans is contingent upon their comprehension of the fact that the probability of getting one's shoes wet is higher when walking alongside a river bank as compared to a financial institution. Machines rarely take the commonsense into account when inferring the meaning of "bank",[4] because they don't have human-like cognition and reasoning abilities by nature.

There are two main technical trends in addressing the task of WSD, namely knowledge-based methods and supervised methods. Knowledge-based WSD utilizes the word relations from knowledge graphs, e.g., WordNet and BabelNet [17] to achieve the disambiguation of word senses. In supervised methods, the WSD task is usually defined as a classification task by word senses. A WSD model is trained with annotated data. Two examples of knowledge-based and supervised WSD are illustrated in Fig. 3. As shown in the figure, a naive strategy of the knowledge-based WSD is that the sense that shares the most relations with the context words is selected as the best-matched one. For supervised WSD systems, the predictive model predicts the potential senses, given the target word and its context words as input. In recent times, the use of knowledge bases has proven advantageous for several modern supervised systems. As a result, there has been a growing trend in integrating knowledge-based and supervised methods to enhance their performance [18].

WSD has been recognized as a crucial module in numerous NLP tasks that heavily rely on word senses, such as sentiment computing, information retrieval, and machine translation. The application of WSD techniques has been demonstrated to be beneficial for these NLP tasks. While prior surveys [19,20] have conducted extensive reviews for WSD, the works discussed in them are outdated. Besides, those works do not link WSD with the linguistic theories and diverse downstream tasks.

2.1. Theoretical research

2.1.1. Distributional semantics

The hypothesis from distributional semantics [21] argued that word meanings can be inferred from word co-occurrences. Words that appear in similar contexts tend to have similar meanings. Such a hypothesis has been the most significant foundation of developing semantic representations in the computational linguistics community, e.g., vector space representations [22–24] and PLMs [8,9]. Based on such a hypothesis, dense semantic vectorial representation research commonly follows a similar training paradigm, e.g., using context words to predict a target word.

Currently, ChatGPT further proves that learning to use words that have appeared before to predict the next possible word can achieve the skills of analogy and reasoning with the help of a very large Transformer [25]-based model.

2.1.2. Selectional preference

Wilks [26] proposed a concept of selectional preference. It is a procedure for representing the meaning structure of natural language. Compared to the "derivational paradigm" of transformational grammar and generative semantics, [26] believed that selectional preference is a more efficient procedure in natural language understanding. It focuses on determining preferences between various possible interpretations of a text, rather than identifying a solitary and unequivocally correct interpretation. Selection preference theory allows more flexibility and nuance in understanding word senses and language. Besides, the theory is computation-friendly. Wilks [26] showed how the procedure could be computed and implemented. The work of Wilks [26] supports that there are multiple possible meanings for a word. The meaning can be defined by the sectional preference of contexts.

2.1.3. Construction semantics

Goldberg and Suttle [27] argued that the meanings of words are frequently derived from larger language units, termed constructions. Constructions consist of a form and a meaning, ranging from single words to full sentences in size. The interpretation of a construction is reliant on both its structure and the situations in which it is employed. Goldberg and Suttle [27] argued that semantic restrictions are better linked with the construction as an entirety rather than with the lexical semantic framework of the verbs. The work of Goldberg and Suttle [27] highlights that the interpretation of meanings of language units can be extended from individual words to constructions. It shows the necessity of defining language units in WSD.

2.1.4. Frame semantics

Fillmore et al. [28] proposed frame semantics that provides a distinct viewpoint on the meanings of words and the principles behind language construction. Frame semantics emphasizes the significance of the surrounding context and encyclopedic knowledge in comprehending word meanings. Petruck [29] explained that a "frame" refers to a collection of concepts interconnected in a manner that understanding any one concept depends on the understanding of the complete system. In frame semantics, the meaning of "cooking" is beyond its dictionary meaning. It also associates with the concept of "food", "cook", "container", and "heating instrument". Frame semantics motivates later ontology research, e.g., FrameNet [30] and FrameNet-based WSD systems, significantly.

2.2. Annotation schemes

For knowledge-based WSD, the data are normally presented as ontology, such as WordNet, FrameNet, and BabelNet, where words and concepts are connected by relations. The relations include hyponyms, hypernyms, holonyms, meronyms, attributes, entailment, etc. An explanation (gloss) and a few example sentences are given for each synset. Synsets of the same Part Of Speech (POS) are connected under some relations independently. However, there exist relations when the basic concept of two words is the same but in a different POS (for example, "propose" and "proposal" were characterized as "derivationally related synsets" in WordNet).

For supervised WSD, a particular word in a given sentence is annotated with a sense ID that corresponds to one of the potential senses in a knowledge base, such as WordNet. A sample of annotation is shown in the next section.

[4] Current methods likely disambiguate word senses by word co-occurrences. However, word co-occurrences are not commonsense.

Table 2
WSD datasets and statistics. SMT, EMEA, KDEdoc, and EUB denote statistical machine translation, European Medicines Agency documents, KDE manual corpus, and the EU bookshop corpus, respectively.

Dataset	Source	# Samples	Reference
SemCor	WordNet	200,000	Miller et al. [31]
MultiSemCor	WordNet, bilingual Collins	51,847	Pianta et al. [32]
Line-hard-serve	WSJ, APHB	4,000	Leacock et al. [33]
Interest	HECTOR	2,369	Bruce and Wiebe [34]
DSO	Brown, WSJ	192,800	Ng and Lee [35]
OMWE	Web	29,165	Chklovski and Pantel [36]
OMSTI	UN documents	1,357,922	Taghipour and Ng [37]
SensEval-2	Unknown	2,282	Edmonds and Cotton [38]
SensEval-3	Editorial, news story, & fiction	1,850	Snyder and Palmer [39]
SemEval2007	Brown, WSJ	455	Pradhan et al. [40]
SemEval2013	SMT workshop	1,644	Navigli et al. [41]
SemEval2015	EMEA, KDEdoc, EUB	1,022	Moro and Navigli [42]

2.3. Datasets

Our surveyed datasets and their statistics can be viewed in Table 2. The biggest manually annotated English corpus currently accessible is SemCor[5] [31]. It has 200K content terms tagged with their related definitions and around 40K phrases. Although SemCor serves as the principal training corpus for WSD, its limited coverage of the English vocabulary for both words and meanings is its most significant drawback. In essence, SemCor merely includes annotations for 22K distinct lexemes in WordNet, the most extensive and commonly employed computerized English dictionary, which corresponds to less than 15% of all words.

To augment the coverage of words, some studies [43] incorporated the English Princeton WordNet Gloss Corpus (WNG),[6] which contains more than 59K WordNet senses, as a complemented data. The WNG is annotated manually or semi-automatically. SemCor and its variations [44,45] lack an acceptable multi-lingual equivalent in the majority of global languages, which limits the scaling capabilities of WSD models beyond English. To address the aforementioned issues, numerous automatic methods for creating multi-lingual sense-annotated data have been developed [46–49]. In an English–Italian parallel corpus known as MultiSemCor [32], senses from the English and Italian versions of WordNet are annotated.

The Line-hard-serve corpus [33] contains 4K samples of the nominal, adjective, and verbal words with sense tags. The data were sourced from Wall Street Journal (WSJ) corpus and the American Printing House for the Blind (APHB) corpus. The Interest corpus [34] contains 2369 occurrences of the term *interest* that have been sense-labeled. The data were sourced from the HECTOR word sense corpus [50]. The Defence Science Organisation (DSO), based in Singapore, created the DSO corpus[7] [35], which contains 192,800 sense-tagged tokens from 191 words from the Brown and WSJ corpora. The Open Mind Word Expert (OMWE) dataset[8] [36] is a corpus of sentences with 288 noun occurrences that were jointly annotated by Web users. One Million Sense-Tagged for Word Sense Disambiguation and Induction (OMSTI)[9] [37] is a semi-automatically annotated WSD dataset with WordNet sense inventory. The data were sourced from MultiUN corpus, which is a collection of United Nation documents.

The SensEval and SemEval datasets are created from the SensEval/SemEval evaluation campaigns. Now, these datasets have been the most widely used benchmarking datasets in WSD. Raganato et al. [51] collected these datasets together[10] and developed a unified evaluation framework for empirical comparison. The statistics of the following datasets are from the collection of Raganato et al. [51]. SensEval-2 [38] used WordNet 1.7 sense inventory, including 2282 sense annotations for nouns, verbs, adverbs and adjectives. SensEval-3 [39] employed WordNet 1.7.1 sense inventory, including 1850 sense annotations. SemEval-2007 Task 17 [40] employed WordNet 2.1 sense inventory, including 455 nominal and verbal sense annotations. SemEval-2013 Task 12 [41] used WordNet 3.0 sense inventory, including 1644 nominal sense annotations. SemEval-2015 Task 13 [42] utilized WordNet 3.0 sense inventory, including 1022 sense annotations. It is worth noting that some of the SemEval tasks are multi-lingual, including SemEval 2013 and 2015, which facilitates multi-lingual WSD.

All of these corpora are annotated using various WordNet sense inventories, with the exception of the Interest corpus (tagged with LDOCE senses) and the Senseval-1 corpus. The Interest corpus and the Senseval-1 corpus were sense-labeled using the HECTOR sense inventories, a lexicon and corpus from a joint Oxford University Press/Digital project [50]. Generally, the data and labels in WSD datasets are organized in the following forms. Then, the task is to identify the sense classes, given contexts, and target words.

```
context: "You perform well in the exam, I will reward you.",
target word: "perform",
pos: "VB",
sense: "3"

context: "She worked in a renowned university for a long time.",
"target word": "university",
"pos": "NN",
"sense": "2"
```

2.4. Knowledge bases

Machine-Readable Dictionaries (MRDs) have been a useful source for WSD due to their structured knowledge and easy access [20]. Dictionaries frequently contain extensive information about the various meanings of a word, as well as illustrative examples of their usage within context. Therefore, dictionaries can serve as valuable knowledge bases for the task of WSD. Additionally, MRDs may provide further information such as synonyms, antonyms, and related words, which can aid in facilitating a better comprehension of a word's meaning. Through the analysis of this information, a system may make more precise determinations about which meaning is most fitting in a given context. There are many electronic dictionaries available for machines to refer to, such as the Longman Dictionary of Contemporary English (LDOCE) [52], the Oxford Dictionary of English (ODE) [53], Collins English Dictionary (CED) [54], and the Oxford Advanced Learner's Dictionary of Current English (OALD) [55] (see Table 3).
WordNet [56] is a sizable, manually curated lexicographic database of English. It is arranged as a network with synsets, or collections of contextual synonyms, as nodes. A synset of synonyms each represents one of a word's senses. Through edges that express lexical-semantic links like meronymies (partof) and hypernymies (is-a), synsets and senses are connected to one another. WordNet additionally offers definitions (glosses) and uses examples for each synset as additional lexical information. English, many WordNets for other languages have been proposed, including languages such as Chinese [57], Arabic [58], Dutch [59], etc.[11]

[5] http://web.eecs.umich.edu/~mihalcea/downloads.html

[6] https://wordnetcode.princeton.edu/glosstag.shtml

[7] https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/QPOJSI

[8] http://web.eecs.umich.edu/~mihalcea/downloads/OMWE/OMWE1.0.English.tar.gz

[9] https://www.comp.nus.edu.sg/~nlp/corpora.html

[10] http://lcl.uniroma1.it/wsdeval/home

[11] See http://globalwordnet.org/resources/wordnets-in-the-world/ for a summary.

Table 3
Useful knowledge bases for WSD. LDOCE means Longman Dictionary of Contemporary English. ODE means Oxford Dictionary of English. CED means Collins English Dictionary. OALD means Oxford Advanced Learner's Dictionary of Current English. Unstructured or structured means the knowledge base contains unstructured or structured lexical knowledge by concepts.

Name	Knowledge	# Entities	Structure
LDOCE 6th ed.	Lexical	230,000	Unstructured
ODE 2022	Lexical	600,000	Unstructured
CED 12th ed.	Lexical	722,000	Unstructured
OALD 8th ed.	Lexical	145,000	Unstructured
WordNet	Lexical	95,600	Graph
FrameNet	Lexical	13,687	Graph
BabelNet	Lexical & Multi-lingual	26,044,643	Graph
SyntagNet	Lexical	78,000	Graph

FrameNet [30] is an English lexical repository that is readable by both humans and machines, established by annotating real-life textual examples that depict the usage of words. It was developed based on the theory of frame semantics, containing 1224 frames (a frame refers to a diagrammatic representation of a scenario encompassing diverse elements such as participants, props, and other conceptual roles), and 13,687 lexical units (lemmas and their PoS) that evoke frames. In FrameNet, the lexical units of a sentence are associated with frame elements. Frame elements are the semantic role of lexical units. For example, given a sentence "I ate an apple this afternoon", "apple" would fill the role of "food" (a frame element).

BabelNet [17] is a multi-lingual dictionary that covers both lexicographic and encyclopedic entries from 520 languages. These entries were created by semi-automatically mapping numerous sites, including WordNet, Multi-lingual WordNet, and Wikipedia. The topology of BabelNet is that of a semantic network, where the nodes are multi-lingual synsets (collections of synonyms that have been lexicalized in several languages), and the edges represent the semantic connections between them.

SyntagNet [60] is a manually developed lexical resource that integrates semantically disambiguated lexical combinations, e.g., noun–verb and noun–noun pairs. The development of SyntagNet involved initially extracting lexical combinations from English Wikipedia and the British National Corpus, which were then subjected to a process of manual disambiguation, based on the WordNet. SyntagNet covers five major languages, e.g., English, German, French, Spanish, and Italian.

2.5. Evaluation metrics

In the WSD task, given a sentence of n words $T = \{x_1, \dots, x_n\}$, the model predicts a sense for each word given the dictionary. Normally, the F1 score is adopted, which is a specialization of the F score when $\alpha = 1$:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{1}$$

Where P denotes precision and R denotes recall:

$$P = \frac{\text{correct predictions}}{\text{total predictions}} \tag{2}$$

$$R = \frac{\text{correct predictions}}{n} \tag{3}$$

The aforementioned metrics do not accurately represent how well systems can produce a level of confidence for a particular sensory choice. Resnik and Yarowsky [61] developed an evaluation criterion that considers the discrepancies between the accurate and selected senses to weigh misclassification mistakes. Therefore, this error will be penalized less severely than coarser sense distinctions if the chosen sense is a fine-grained distinction of the true sense. There have been evaluation metrics for even more precise measurements, including the Receiver Operation Characteristic (ROC) [62]. However, compared with traditional metrics such as precision, recall, and F1, these metrics are not frequently utilized.

2.6. Annotation tools

LX-SenseAnnotator[12] [63] provides a user interface for manually annotating word senses. The software has the capability to process lexical data in any language, on the condition that the data is compliant with the format of Princeton WordNet. Human annotators can view the pre-processed text in three different modes, including the source text, sense-annotated text, and raw text, which can be switched between by using a tab widget. The source text mode displays the original text along with all tags, while the sense-annotated text mode displays the same text but with newly added sense tags. This allows the annotator to monitor the output file continually. Annotators can view the sense options in real time when annotating the sense for a word.

LexTag[13] is another useful tool for WSD. The annotation interface provided is characterized by its user-friendly nature, facilitating users in the annotation of various textual elements such as terms, sentences, and documents. This annotation process involves attributing meanings drawn from pre-existing knowledge graphs and dictionaries, encompassing reputable sources like WordNet, Wiktionary, and WordAtlas. LexTag has been used to create a recent 10-language parallel dataset ELEXIS-WSD 1.0.[14]

2.7. Methods

2.7.1. Knowledge-based WSD

Knowledge-based WSD utilizes knowledge bases to disambiguate word senses. Compared with supervised WSD, this class of WSD methods achieves lower performance but better data efficiency. In knowledge-based WSD, there are essentially two research streams.

A. Semantic Space Matching

One stream of the knowledge-based WSD is to look for overlaps or similarities between the context of a term whose sense needs to be disambiguated and its sense representation, such as the definition of a potential sense and its associated sense that was retrieved from a knowledge base. The predicted sense is considered to be the sense that is the closest.

Lesk [64] is a naive knowledge-based WSD algorithm that looks for terms that are similar to the target word in the context of each sense. The approach aimed to enumerate the intersections among lexicon definitions of the diverse connotations of every target word contained within a given sentence. Banerjee et al. [65] proposed an advanced version of the Lesk, which also includes the definition of related senses, where the standard term frequency-inverse document frequency method is employed for word weighting. Another improved version of Lesk [66] includes word embedding for better analysis, which improves the accuracy of determining how close the definition and context of the target word are. SREF$_{KB}$ [18] is a state-of-the-art (SOTA) WSD system. It is a vector-based technique that disambiguates word senses by using sense embeddings and contextualized word representations. It applied BERT to represent WordNet instances and definitions, as well as the automatically obtained contexts from the Web.

B. Graph-based Matching

The other stream of the knowledge-based WSD creates a graph using the given context and connections that have been retrieved from knowledge bases. Here, the synsets and the relationships between them are seen as the nodes and edges, respectively. The senses are then disambiguated based on the constructed graphs. A variety of graph-based techniques, such as Latent Dirichlet Allocation (LDA) [67], PageRank [68], Random Walks [69], Clique Approximation [70], Game Theory [71], etc., are used to disambiguate the meaning of a given word using the created graph.

[12] http://nlx.di.fc.ul.pt/tools.html
[13] https://babelscape.com/lextag
[14] https://www.clarin.si/repository/xmlui/handle/11356/1674

Agirre and Soroa [72] presented a graph-based unsupervised WSD system that employs random walk over a WordNet semantic network. They employed a customized version of the Page Rank algorithm [73]. The technique leverages the inherent structural properties of the graph that underlies a specific lexical knowledge base, and shows the capability of the algorithm to identify global optima for WSD, based on the relations among entities. Agirre et al. [69] evaluated this algorithm with new datasets and variations of the algorithm to prove its effectiveness. Navigli and Lapata [74] also introduced a graph-based unsupervised model for WSD, which analyzed the connectivity of graph structures to identify the most pertinent word senses. A graph is constructed to represent all possible interpretations of the word sequence, where nodes represent word senses and edges represent sense dependencies. The model assessed the graph structure to determine the significance of each node, thus finding the most crucial node for each word. Babelfy [70] is also a graph-based WSD method that uses random walk to identify relationships between synsets. It used BabelNet [17] and performed random walks with Restart [75]. In addition, it incorporated the entire document at the time of disambiguation. The candidate disambiguation is upon automatically developed semantic interpretation graph which used a graph structure to represent various possible interpretations of input text. SyntagRank [76] is a high-scoring knowledge-based WSD algorithm. It is an entirely graph-based algorithm that uses the Personalized PageRank algorithm to incorporate WordNet (for English), BabelNet (for non-English) and SyntagNet. SyntagRank is generally considered a stronger method than SREF$_{KB}$. BabelNet enabled SyntagRank to improve its ability to scale across a wide range of languages, whereas SREF$_{KB}$ has only been evaluated in English.

2.7.2. Supervised WSD

Currently, supervised approaches, especially deep learning-based supervised learning approaches, have become mainstream in the WSD community. Earlier deep learning-based approaches focused on architectures where WSD was defined as token classification over WordNet senses [77]. Even though they performed well, these structures showed a lot of flaws, particularly when it came to predicting uncommon and invisible senses. To address these issues, numerous works began to supplement the training data by utilizing various lexical knowledge, such as sense definitions [78,79], semantic relations [80,81], and data generated via novel generative methods [82]. In this section, we review the representative works in supervised WSD.

A. Data-Driven Machine Learning Approaches

Data-driven machine learning approaches refer to methodologies and techniques in which the design, training, and optimization of traditional machine learning algorithms, heavily rely on large amounts of data. In these approaches, the model's ability to generalize patterns and make predictions is learned directly from the provided data, rather than being explicitly programmed by humans. In the early days, classic machine learning approaches with handcrafted features were frequently used for WSD. Singh et al. [83] employed 5-g and position features, and a decision tree algorithm to represent classification rules in a tree structure where the training dataset is recursively partitioned. Each leaf node indicates the meaning of a word. They developed a dataset, containing 672 Manipuri sentences to test their method. The sentences were sourced from a local newspaper, termed "The Sangai Express". O'Hara et al. [84] proposed a class-based collocation method that integrates diverse linguistic features in a decision tree algorithm. For the collocation, three distinct word relatedness scores are used: the first is based on WordNet hypernym relations; the second is based on cluster-based word similarity classes; and the third is based on dictionary definition analysis. The authors also utilized PoS and word form features. The It Makes Sense (IMS) WSD system [85] used a Support Vector Machine (SVM) classifier. Different positional and linguistic features were considered, including nearby words, nearby words' PoS tags, and nearby collocations. Later, word embeddings became important features in WSD. Taghipour and Ng [37]; Rothe and Schütze [86]; Iacobacci et al. [87] used IMS as the base model to examine word embeddings. Iacobacci et al. [87] offered many approaches where different word embeddings were applied as features to test how many parameters impact the effectiveness of a WSD system. The authors found that word2vec [23] which was trained with OMSTI can yield the strongest results on the three examined all-word WSD tasks.

B. Data-Driven Neural Approaches

More recently, neural approaches started to be used. Data-driven neural approaches refer to methodologies and techniques that utilize neural networks and supervised learning to learn patterns and representations directly from data.

Popov [88] proposed to use BiLSTM [89], GloVe word embeddings, and word2vec lemma embeddings. Yuan et al. [90] suggested another LSTM-based word sense disambiguation approach that was trained in a semi-supervised fashion. The semi-supervised learning was achieved by employing label propagation [91] to assign labels to unannotated sentences by assessing their similarity to labeled ones. The best performance on the SensEval-2 dataset can be observed from the model that was semi-supervision-trained with OMSTI and 1000 additional unlabeled sentences. Additionally, [92] looked more closely at how many elements affect its performance, and several intriguing conclusions were drawn. The initial point to highlight is that achieving strong WSD performance does not necessitate an exceedingly large unannotated dataset. Furthermore, this method provides a more evenly-distributed sense assignment in comparison to prior approaches, as evidenced by its relatively strong performance on infrequent cases. Additionally, it is worth noting that the limited sense coverage of the annotated dataset may serve as an upper limit on overall performance.

With the development of self-attention-based neural architectures and their capacity to extract sophisticated language information [25], the use of transformer-based architectures in fully supervised WSD systems is becoming more and more popular. The WSD task is usually fine-tuned on a pre-trained transformer model, which is a popular strategy. The task-specific inputs are given to the pre-trained model, which is then further trained across a number of epochs with the task-specific objective. Likewise, in recent token classification models for WSD, the contextualized representations are usually generated by a pre-trained model and then fed to either a feedforward network [93] or a stack of Transformer layers [94]. These methods outperform earlier randomly initialized models [95]. Hadiwinoto et al. [93] tested different pooling strategies of BERT, e.g., last layer projection, weighted sum of hidden layers, and Gated Linear Unit [96]. The best performance on SensEval-2 is given by the strategy of the weighted sum of hidden layers, accounting for 76.4% F1. Bevilacqua and Navigli [94] proposed a bi-directional Transformer that explicitly attends to past and future information. This model achieved 75.7% F1 on SensEval-2 by training with the combination of SemCor and WordNet's Tagged Glosses.[15] It is worth noting that, the categorical cross-entropy, which is frequently utilized for training, limits the performances. In reality, it has been demonstrated that the binary cross-entropy loss performs better [81] because it enables the consideration of many annotations for a single instance in the training set as opposed to the use of a single ground-truth sense alone. In the above-mentioned approaches, each sense is assumed to be a unique class, and the classification architecture is limited to the information provided by the training corpus.

2.7.3. Knowledge-augmented supervised WSD

The edges that connect the senses and synsets are a valuable source of knowledge that augments the annotated data. Traditionally, graph knowledge-based systems, such as those based on Personalized PageRank [76], have taken advantage of this information. Moreover, utilizing

[15] https://wordnetcode.princeton.edu/glosstag.shtml

WordNet as a graph has benefited many modern supervised systems. Thus, formally, knowledge-augmented supervised WSD is defined as a methodology that combines traditional supervised machine learning techniques with external knowledge resources to improve the accuracy and performance of word sense disambiguation.

Wang and Wang [18] used WordNet hypernymy and hyponymy relations to devise a try-again mechanism that refines the prediction of the WSD model. The SemCor corpus was utilized to acquire a supervised sense embedding for every annotated sense in their supervised method (SREF$_{Sup}$). Vial et al. [43] reduced the number of output classes by mapping each sense to an ancestor in the WordNet taxonomy, then yielding a smaller but robust sense vocabulary. The authors used BERT contextualized embeddings. By training with SemCor and WordNet gloss corpora, the model achieved 79.7% F1 on SensEval-2. Different variations also achieve outstanding performance on diverse WSD datasets.

Loureiro and Jorge [97] created representations for those senses not appearing in SemCor by using the averaged neighbor embeddings in the WordNet. The token-tagger models EWISE [78] and EWISER [80] both leveraged the WordNet graph structure to train the gloss embedding offline, where EWISER demonstrated how the WordNet entire graph feature can be directly extracted. EWISE used ConvE [98] to obtain graph embeddings. Conia and Navigli [81] provided a new technique to use the same edge information by replacing the adjacency matrix multiplication with a binary cross-entropy loss where other senses connected to the gold sense are also taken into account. The edge information was obtained from WordNet. In general, edge information is increasingly used in supervised WSD, gradually blending with knowledge-based techniques. However, it can only be conveniently utilized by token classification procedures, whereas its incorporation into sequence classification techniques has not yet been researched.

It has also been extensively studied how to use sense definitions as an additional source for supervised WSD apart from the traditional data annotations. It considerably increased the scalability of a model on the senses that are underrepresented in the training corpus. Huang et al. [99] argued that WSD has traditionally been approached as a binary classification task, whereby a model must accurately decide if the sense of a given word in context aligns with one of its potential meanings in a sense inventory, based on the provided definition. Define the WSD task as a sentence-pair classification task, where the WordNet gloss of a target word is concatenated after an input sentence. Blevins and Zettlemoyer [79] used a bi-encoder to project both words in context and WordNet glosses in a common vector space. Disambiguation is then carried out by determining the gloss that is most similar to the target word. Glosses are employed similarly by more advanced techniques like SensEmBERT [100], ARES [101], and SREF [18]. They used quite different approaches to find new contexts automatically in order to develop the supervised portion of the sense embedding. ARES achieved 78.0% F1 on the SensEval-2 dataset by utilizing collocational relations between senses to get novel example sentences from websites. SensEmBERT leveraged BabelNet and Wikipedia explanations, achieving significant improvements on nominal WSD tasks over 5 major datasets. Barba et al. [102] proposed to solve WSD as a text extraction problem where, given a word in context and all of its potential glosses, models extract the definition that best matches the term under consideration. The authors demonstrated the advantages of their approach in that it does not require huge output vocabularies and enables models to take into account both the input context and all meanings of the target word simultaneously. By using sparse coding, [103] has demonstrated that it is also possible to make existing sense embeddings sparse. All of these methods handle each word independently of the others when disambiguating multiple words that co-occur in the same context. Thus, a word's explicit meaning is neither taken into account during word disambiguation nor does it have an impact on the disambiguation of surrounding words.

2.8. Downstream applications

2.8.1. Sentiment computing

WSD has been applied in many Sentiment Analysis (SA) works to improve accuracy and explainability. Farooq et al. [104] proposed a WSD framework to enhance the performance of sentiment analysis. To determine the orientation of opinions related to product attributes in a particular field, a lexical dictionary comprising various word senses is developed. The process involves extracting relevant features from product reviews and identifying opinion-bearing texts, followed by the extraction of words used to describe the features and their contexts to form seed words. These seed words, which consist of adjectives, nouns, verbs, and adverbs, are manually annotated with their respective polarities, and their coverage is extended by retrieving their synonyms and antonyms. WSD was utilized to identify the sentiment-orientated senses, such as the positive, negative, or neutral senses of a word in a sentence, because a word may have different sentiment polarities by taking different senses in different contexts.

Nassirtoussi et al. [105] offered a novel approach to forecast intraday directional movements of the EUR/USD exchange rates based on news headline text mining in an effort to address semantic and sentiment components of text-mining. They evaluated news headlines semantically and emotionally using the lexicons, e.g., WordNet and SentiWordNet [106]. SentiWordNet is a publicly accessible lexical resource designed for sentiment analysis that allocates a positivity score, negativity score, and objectivity score to each synset within WordNet. Nassirtoussi et al. [105] found that both positive and negative emotions may influence the market in the same way. WSD worked as a technique to abstract semantic information in their framework. Thus, it enhances the feature representations and explainability in their downstream task modeling. SentiWordNet has served as a basis for various sentiment analysis models. In the work of Ohana and Tierney [107], the feasibility of using the emotional scores of SentiWordNet to automatically classify the sentiment of movie reviews was examined. Other applications, e.g., business opinion mining [108], article emotion classification [109], word-of-mouth sentiment classification [110,111] also showed that SentiWordNet as a semantic feature enhancement knowledge base can deliver accuracy gains and model insights in sentiment analysis tasks.

2.8.2. Information retrieval

The impacts of using WSD for information retrieval have been examined in many works. Krovetz and Croft [112] disambiguated word senses for terms in queries and documents to examine how ambiguous word senses impact information retrieval performance. The researchers arrived at the conclusion that the advantages of WSD in information retrieval are marginal. This is due to the fact that query words have uneven sense distributions. The impact of collocation from other query terms already plays a role in disambiguation. WSD was used as a parser to study this task. However, the findings from [113] are different. They examined the impact of improper disambiguation using SemCor. By accurately modeling documents and queries together with synsets, they achieved notable gains (synonym sets). Additionally, their study demonstrated that WSD with an error rate of 40%–50% may still enhance IR performance when used with the synset representation, which incorporated synonym information. Gonzalo et al. [114],Stokoe et al. [115] further confirmed the significance of WSD to information retrieval. Gonzalo et al. [114] also found that PoS information has a lower utility for information retrieval. Based on artificially creating word ambiguity, [116] employed pseudo words to explore the effects of sense ambiguity on information retrieval. They came to the conclusion that the high accuracy of WSD is a crucial condition to accomplish progress. Blloshmi et al. [117] introduced an innovative approach to multi-lingual query expansion by integrating WSD, which augments the query with sense definitions as supplementary semantic information in multi-lingual neural ranking-based IR.

The results demonstrated the advantages of WSD in improving contextualized queries, resulting in a more accurate document-matching process and retrieving more relevant documents. Kim et al. [118] labeled words with 25 root meanings of nouns rather than utilizing fine-grained sense inventories of WordNet. Their retrieval technique preserved the stem-based index and changed the word weight in a document in accordance with the degree to which it matched the query's sense. They credited their coarse-grained, reliable, and adaptable sense tagging system with the improvement on TREC collections. The detrimental effects of disambiguation mistakes are somewhat mitigated by the addition of senses to the conventional stem-based index.

2.8.3. Machine translation

The challenge of ambiguous word senses poses a significant barrier to the development of an efficient machine translator. As a result, a number of researchers have turned their attention to exploring WSD for machine translation. Some works tried to establish datasets to quantify the WSD capacity of machine translation systems. Rios Gonzales et al. [119] proposed a test set of 6700 lexical ambiguities for German–French and 7200 for German–English. They discovered that WSD remains a difficult challenge for neural machine translation, especially for uncommon word senses, even with 70% of lexical ambiguities properly resolved. Campolungo et al. [120] proposed a benchmark dataset that aims at measuring WSD biases in Machine Translation in five language combinations. They also agreed that SOTA systems still exhibited notable constraints when confronted with less common word senses. Incorporating sense labels and lexical chains leads to enhanced performance of Neural Machine Translation (NMT) models, particularly with regard to infrequent word senses. Raganato et al. [121] proposed MUCOW, a multi-lingual contrastive test set automatically created from word-aligned parallel corpora and the comprehensive multi-lingual sense inventory of BabelNet. MUCOW spans 16 language pairs and contains more than 200,000 contrastive sentence pairs. The researchers thoroughly evaluated the effectiveness of the ambiguous lexicons and the resulting test suite by utilizing pre-trained NMT models and analyzing all submissions across nine language pairs from the WMT19 news shared translation task.

Some works analyzed the internal representations to understand the disambiguation process in machine translation systems. Marvin and Koehn [122] examined the extent to which ambiguous word senses could be decoded through the use of word embeddings in relation to deeper layers of the NMT encoder, which were believed to represent words with contextual information. In line with prior research, they discovered that the NMT system frequently mistranslated ambiguous terms. Tang et al. [123] trained a classifier to determine if a translation is accurate given the representation of an ambiguous noun. The fact that encoder hidden states performed much better than word embeddings suggests that encoders are able to appropriately encode important data for disambiguation into hidden states. Liu et al. [124] discovered that an increase in the number of senses associated with each word results in a decline in the performance of word-level translation. The root of the issue may be the mapping of each word to similar word vectors, regardless of its context. They proposed to integrate techniques from neural WSD systems into an NMT system to address this issue.

2.9. Summary

WSD as a computational linguistics task most closely related to lexical semantics research, has won extensive discussions among researchers from different fields. Linguists came up with important hypotheses to guide the modeling of word senses. We have observed that some hypotheses have been well grounded in NLP, e.g., learning and representing word meanings with their contexts and word co-occurrences. However, we also observe some important linguistic arguments were rarely studied in the computational linguistic domain, e.g., defining the scope of linguistic units for WSD and integrating relevant concepts (frames) for word sense representations. The development of WSD datasets has greatly ignited the research enthusiasm of scholars in WSD. However, we also observed that the computational research on WSD is also limited by these well-defined datasets because WSD datasets generally follow a very similar labeling paradigm. Relevant linguistic studies have shown broader possibilities in WSD. Finally, we find that many of WSD modeling techniques do not link well with downstream applications. The research of WSD methods has intersections with downstream applications, whereas they cannot well cover the needs of downstream tasks. This also shows that the research opportunities in WSD can be largely extended besides word sense classification.

2.9.1. Technical trends

Table 4 shows the technical trends of WSD methods. As seen in the table, earlier approaches likely used knowledge-based and supervised approaches. WordNet and BabelNet are useful knowledge bases that were frequently used by knowledge-based methods. Word embeddings, pre-trained language models, and linguistic features, e.g., PoS tags and semantic relatedness were frequently used by supervised methods. For old pure knowledge-based methods, the PageRank framework was likely used, because many knowledge bases are represented as graphs. PageRank is an algorithm used in graph computation to measure the importance of nodes in a graph. Classical machine learning techniques, e.g., Decision Tree, SVM, LSTM, and Transformers were commonly used by supervised WSD methods. Supervised learning algorithms demonstrate superior performance in comparison to knowledge-based approaches. Nevertheless, it is not always reasonable to assume the availability of substantial training datasets for different areas, languages, and activities. Ng [125] predicted that a corpus of around 3.2 million sense-tagged words would be necessary in order to produce a high-accuracy, wide-coverage disambiguation system. The creation of such a training corpus requires an estimated 27 person-years of labor. The accuracy of supervised systems might be greatly improved above the SOTA methods with such a resource. However, the success of this hypothesis is at the cost of huge resource consumption.

We observe more hybrid approaches that leverage knowledge bases in a supervised learning fashion in recent years. This is because researchers have observed the limitations of typical supervised WSD in processing rare or unseen cases. Knowledge bases provide additional information to support the learning of unseen cases. Knowledge bases provide additional knowledge for the languages whose annotated data are scarce. In this case, multi-lingual knowledge bases can enhance the representations of word senses in a new domain. As a result, we can observe the accuracy of the hybrid approaches surpasses the pure knowledge-based or supervised approaches.

Most existing WSD datasets define the task as a word sense classification task. Then, the following methodology research upon the datasets focused on improving the accuracy of mapping the sense of a word to its dictionary sense class. However, should the research on WSD be limited to word sense classification? We have observed that many knowledge-based systems used existing knowledge bases to conduct word sense classification tasks. They have realized the importance of developing an effective knowledge base for WSD. However, it is rare to see that WSD research tries to improve the construction of knowledge bases according to the effectiveness of word sense classification. On the other hand, the meaning of WSD is much larger than detecting the definition of words in a dictionary. Mapping a word to a sense in a dictionary is just an aspect of WSD. Previous works rarely studied what is an appropriate linguistic unit for WSD; what concepts are associated with a word sense in a context. These are very interesting research topics from linguistic and cognitive aspects. However, these topics were not well studied in the computational WSD community.

Table 4
A summary of representative WSD techniques. Knwl denotes knowledge-based methods. Sup. denotes supervised methods. KB denotes knowledge bases. WN denotes WordNet. BN denotes BabelNet. DSM denotes Distributional Semantics Models. Prob. denotes probability. SE2013-EN denotes the SemEval2013 English WSD task. PMI denotes Pointwise Mutual Information.

Task	Reference	Tech	Feature and KB.	Framework	Dataset	Score	Metric
Knwl	Lesk [64]	Prob.	Statistics, OALD	Count def. overlaps	–	–	–
	Banerjee et al. [65]	ML	Emb., WN	Score function	SensEval-2	34.60%	F1
	Navigli and Lapata [74]	Graph	Sense graph, WN	Connectivity measures	SemCor	31.80%	F1
	Basile et al. [66]	Prob.	Emb., BN	DSM	SE2013-EN	71.50%	F1
	Wang and Wang [18]$_{KB}$	DL	BERT, WN	Vector represent.	SensEval-2	72.70%	F1
	Agirre and Soroa [72]	Graph	WN	PageRank	SensEval-2	58.60%	Recall
	Moro et al. [70]	Graph	Sem. graph, BN	PageRank	SE2013-EN	69.20%	F1
	Scozzafava et al. [76]	Graph	WN, SN	PageRank	SensEval-2	71.60%	F1
Sup.	Singh et al. [83]	ML	5-g, position	Decision Tree	Manipuri	71.75%	Acc
	O'Hara et al. [84]	ML	Relatedness scores	Decision tree	SensEval-3	65.90%	F1
	Zhong and Ng [85]	ML	Position, PoS	SVM	SensEval-2	68.20%	F1
	Iacobacci et al. [87]	ML	Emb., position, PoS	SVM	SensEval-2	68.30%	F1
	Popov [88]	DL	Emb.	BiLSTM	SensEval-2	70.11%	Acc
	Yuan et al. [90]	DL	Emb., label propag.	LSTM	SensEval-2	74.40%	F1
	Le et al. [92]	DL	Emb.	LSTM	SensEval-2	72.00%	F1
	Hadiwinoto et al. [93]	DL	BERT	Transformer	SensEval-2	76.40%	F1
	Bevilacqua and Navigli [94]	DL	Emb.	BiTransformer	SensEval-2	75.70%	F1
Knwl + Sup.	Wang and Wang [18]$_{Sup}$	DL	BERT, WN	Vector represent.	SensEval-2	78.60%	F1
	Vial et al. [43]	DL	BERT, WN	Transformer	SensEval-2	79.70%	F1
	Loureiro and Jorge [97]	DL	BERT, WN	Transformer	SensEval-2	76.30%	F1
	Kumar et al. [78]	DL	Graph emb., emb., WN	BiLSTM, Att. ConvE	SensEval-2	73.80%	F1
	Bevilacqua and Navigli [80]	DL	BERT, WN	Trans., Struct. logit	5 datasets	80.80%	F1
	Conia and Navigli [81]	DL	BERT, WN	Transformer	SensEval-2	78.40%	F1
	Huang et al. [99]	DL	BERT, WN	Transformer, sentence-pair classification	SensEval-2	77.70%	F1
	Blevins and Zettlemoyer [79]	DL	BERT, WN	Trasformer, Score func.	SensEval-2	79.40%	F1
	Scarlini et al. [100]	DL	BERT, BN, Wiki	Transformer, Context retrieval	Nouns of 5 datasets	80.40%	F1
	Scarlini et al. [101]	DL	BERT, WN, SN	Transformer, Context retrieval	SensEval-2	78.00%	F1
	Barba et al. [102]	DL	BERT, WN	Transformer, Extractive sense learning	SensEval-2	81.70%	F1
	Berend [103]	DL	BERT, WN	Transformer, sparse coding, PMI	SensEval-2	79.60%	F1

Table 5
A summary of the representative applications of WSD in downstream tasks. ✓ denotes the role of WSD in a downstream task.

Reference	Downstream task	Feature	Parser	Explainability
Farooq et al. [104]	Sentiment computing	✓		
Nassirtoussi et al. [105]	Sentiment computing	✓		✓
Ohana and Tierney [107]	Sentiment computing	✓	✓	
Saggion and Funk [108]	Sentiment computing	✓	✓	✓
Devitt and Ahmad [109]	Sentiment computing	✓		✓
Hung and Lin [110]	Sentiment computing	✓	✓	
Hung and Chen [111]	Sentiment computing	✓	✓	✓
Krovetz and Croft [112]	Information retrieval		✓	
Gonzalo et al. [113]	Information retrieval			✓
Gonzalo et al. [114]	Information retrieval			✓
Sanderson [116]	Information retrieval			✓
Stokoe et al. [115]	Information retrieval			✓
Kim et al. [118]	Information retrieval	✓	✓	
Blloshmi et al. [117]	Information retrieval	✓		✓
Rios Gonzales et al. [119]	Machine translation	✓		
Raganato et al. [121]	Machine translation	✓		
Marvin and Koehn [122]	Machine translation		✓	✓
Tang et al. [123]	Machine translation	✓		✓
Liu et al. [124]	Machine translation	✓		

2.9.2. Application trends

The WSD task was commonly defined as a word sense classification task. However, we observe that classifying words by sense classes is not the only need for downstream NLP tasks (see Table 5). There are three main tasks that are strongly related to WSD, e.g., sentiment computing, information retrieval, and machine translation in our survey. One of the roles of WSD on the three tasks is to deliver or enhance features to gain improvements on the three tasks. On the other hand, we also observe many downstream works used WSD techniques as a parser to obtain words with different levels of word sense ambiguity or used WSD to gain insights into their model behaviors to improve the explainability of a study. In these cases, defining WSD as a sense classification task may be sub-optimal for downstream applications.

WSD has a huge potential in NLP research. For example, disambiguating word senses in a large corpus can lead to a deeper understanding of language usage patterns and the semantic relationships between words. WSD is also a significant component in semantic explainable AI, because it helps researchers better understand the decision-making process of a model on the semantic level. Researchers can develop a more transparent and trustworthy model by explaining word senses in contexts. As a feature generator, a WSD may be more effective if it can generate contextualized word meanings in natural language, rather than predict a sense class that maps to a predefined gloss in a dictionary. However, research in these fields is rare in the WSD community.

Finally, according to [20], the lack of end-to-end applications that utilize WSD can be attributed to the insufficient accuracy of current WSD systems. This suggests that in the future, more precise WSD systems may be developed, which could potentially enable the use of more semantics-dependent applications.

2.9.3. Future works

As argued before, the task of WSD can be broader than the current word sense classification task setup from either the theoretical research side or the downstream application side. Besides, the improvements in WSD accuracy can also attract more downstream applications. Thus, we come up with the following future work suggestions.

Extending the form of WSD. WSD can have different learning forms, besides word sense classification, e.g., paraphrasing an ambiguous word into a less ambiguous one [126,127], generating contextualized word senses in natural language. Such an extension may have significance in downstream applications. From the perspective of linguistic and cognitive research, studying how to define a language unit to better disambiguate word senses, or studying how to link a word to its associated concepts in a context can also improve the significance of WSD in the era of LLM-based NLP. Future works may study how to define the task of WSD to better support the research in different disciplines.

Rethinking existing knowledge bases by WSD. Most of the existing knowledge bases were developed according to human-defined ontologies and word senses. These knowledge bases have been considered as an important resource for many knowledge-based systems. Although the knowledge bases have been used on different tasks, few works analyzed the weakness of the ontologies. Future WSD-related research may try to improve the knowledge bases by rethinking the sense definition, concept node connections, and coverage, rather than simply developing models to enhance the learning ability on a specific task.

Multi-lingual WSD. Most of the semantic representations are learned from monolingual corpora. As a result, the semantic representations are different between different languages. However, the disambiguation of meanings is not characterized by languages [128]. It will significantly improve multi-lingual semantic research if WSD research can break down language barriers from a cognitive perspective. As argued by frame semantics [28], the meaning of a word is beyond its dictionary definitions. It also associates with the concepts, interconnected with the word. Representing word senses by concepts may achieve a more robust multi-lingual WSD.

Learning WSD as a pre-training task. Recent years witness great success of PLMs in various domains. The existing PLMs followed the same hypothesis that the sense of a word can be learned from its associated context. However, there has not been a PLM that explicitly disambiguates word senses to enhance the learning of semantic representations. Naively learning the semantic representation of a target word by its associated context words cannot learn the conceptual association of the target word. For example, many words can associate with the word “apple”. How can we know an apple as fruit is red or green, sweet, tree-growing, nutritious, etc? As an electronic device, Apple is associated with an operating system, a circuit board, a brand, etc. Disambiguating word senses before pre-training may build such connections between concepts.

Fusing WSD with other tasks. As [129] argued, WSD can also be integrated with an entity linking task [70], where the model predicts associated entities to help WSD systems explore the related glosses and relations. Related fusion works also include fusing WSD for Sentiment Analysis [104], Information Retrieval [117] and Machine Translation [120]. The future study of WSD can be grounded on an end task so that the end task can more effectively benefit from the fusion of a WSD model.

3. Anaphora resolution

In computational linguistics, Ruslan Mitkov defined anaphora as a *phenomena of pointing back a previously mentioned item in the text* [130]. The pointing back phrase is called an *anaphor* while the previously mentioned item is called an *antecedent*.

The concept of anaphora should not be confused with co-reference. On the one hand, either anaphora or cataphora (e.g., the phenomena of pointing ahead to a subsequently mentioned item) could be a kind of co-reference. On the other hand, an anaphor and its antecedent are not always co-referential. By definition, the difference between anaphora and co-reference is that anaphora does not require *identify-of-reference* while co-reference requires. In other words, anaphora may describe a relation between expressions that do not have the same referent. For example, in sentence (2), the anaphor “one” has the same sense as its antecedent “a dog”, but they do not refer to the same dog.

(2) Jack has a dog and Mary also has **one**.

Building on this, in relation to anaphora, both anaphor and its antecedent are not necessarily referring expressions. For instance, an anaphor can be a verb (henceforth, verb anaphora). In the following example from [131],

(3) When Manchester United swooped to lure Ron Atkinson away from the Albion, it was inevitable that his midfield prodigy would **follow**, and in 1981 he **did**.

the anaphor “did” is a verb, having an antecedent “follow”. Another example is the *bound anaphora* where the antecedent is a quantified expression [132]:

(4) Each manager exploits the secretary who works for “him”.

The anaphor “him” refers to the quantified expression “each manager”. Since antecedents in both above two examples are not referring expressions, neither of them is a co-reference.

Given the definition of anaphora, the task of anaphora resolution is to identify the antecedent of an anaphor. In this survey, we decided to merely focus on anaphora resolution (rather than co-reference resolution) because, on the one hand, most semantic processing tasks only require identifying antecedents. On the other hand, we are not only interested in referring to noun phrases but also other phrases that an anaphor can refer to (e.g., verb phrases and quantified expressions; see the discussion above).

It is worth noting that there have been reviews in the past 20 years about AR/CR from either computer scientists [133,134] or linguists [13,130]. In this survey, our objective is to establish a connection between AR techniques across theoretical research and practical applications.

3.1. Theoretical research

3.1.1. Constraints

When human beings resolute co-reference, there are semantic and syntactic constraints. As for the semantic constraints, agreements such as gender and number agreements are the strongest type [135]. However, most recently, agreement mismatch problems (especially for gender agreements) have been becoming more frequent since more people have started to use plural pronouns to avoid gender bias.

As for syntactic constraints, according to the binding theory [136], in the sentence (a) of the following example, “John” cannot co-refer with “him” while in the sentence (b) “John” can.

(5) a. John likes him.
b. John likes him in the mirror.

3.1.2. Centering theory

Centering Theory [137–139] was introduced as a model of *local coherence*[16] based on the idea of *center of attention*. The theory assumes that, during the production or comprehension of a discourse, the discourse participant’s attention is often centered on a set of entities (a subset of all entities in the discourse) and such an *attentional state* evolves dynamically. It models transitions of the attentional state and defines three types of transitions: CONTINUE, RETAIN, and SHIFT. For each utterance, the transition is decided by its backward-looking center (defined as the most salient entity in the previous utterance that is also realized in the current utterance and denoted as C_b) as well as forward-looking center (defined as the most salient entity in the current utterance and denoted as C_f). Consider the following discourse adopted from [140]:

(6) a. Terry really gets angry sometimes.
b. Yesterday was a beautiful day and he was excited about trying out his new sailboat. [C_b = Terry, C_f = Terry]
c. He wanted Tony to join him on a sailing expedition, and left him a message on his answering machine. [C_b = Terry, C_f = Terry]
d. Tony called him at 6AM the next morning. [C_b = Terry, C_f = Tony]
e. Tony was furious with him for being woken up so early. [C_b = Tony, C_f = Tony]

[16] Instead of focusing on the whole discourse, centering theory focuses only on the *discourse segment*.

where we annotate each utterance with its backward-looking and forward-looking centers. The transition from utterance a to b is a CONTINUE as both backward-looking and forward-looking centers are unchanged. The next one is a RETAIN transition since although the most salient entity changes (i.e., C_f), the forward-looking center stays the same, whereas the transition from utterance d to e is a SHIFT transition because of the change of backward-looking transition. Intuitively, a discourse with more CONTINUE transitions is more coherent than the one with more SHIFT transitions.

Though Centering Theory is not a theory of Anaphora Resolution, Anaphora Resolution can directly benefit from modeling transitions, which provides certain information about the preference for the referents of pronouns (e.g., in a coherent segment, centers co-refer; see [141] for more discussion about the relation between Centering Theory and Anaphora Resolution).

3.1.3. Discourse salience

A prominent strand of work in psycholinguistics investigates how human beings use anaphora. A referent is more likely to be realized as a pronoun if it is salient in a given discourse [142] (aka. *discourse salience*). Discourse salience is thought to be influenced by various factors, including givenness [143,144], grammatical role [145,146], recency [142,147], syntactic parallelism [147,148], and many other factors. Similar to Centering Theory, most research on discourse salience is about the production of anaphora [149–152], but it also provides insights about an antecedent's relative likelihood for a given anaphor in a given discourse. In this sense, it is plausible to use the aforementioned factors as features to rank candidate antecedents of an anaphor [153,154].

3.1.4. Coolness

Huang [155] classified human languages into cool languages and hot languages. If a language is "cooler" than another language, then understanding a sentence in that language relies more on context (see [156–158] for computational investigations of the theory of Coolness). The evidence that [155] identified is about the differences between the use of anaphora. Specifically, cool languages (e.g., Mandarin) make liberal use of zero pronouns. Take the following conversation as an example:

(7) a. 你今天看见比尔了吗？(Did you see Bill today?)
b. *pro*看见*pro*了。(*I* saw *him*.)

where a *pro* represents a zero pronoun[17] (ZP). The first ZP refers to one of the speakers while the second ZP refers to Bill. ZPs of this kind are called Anaphoric ZPs (AZPs). In addition to Mandarin, a number of other languages (i.e., cool languages) also allow ZPs, including examples like Japanese, Arabic, and Korean. The current theory suggests that the anaphora resolution of cool languages should also take AZPs into consideration, namely AZP resolution [159].

3.2. Annotation schemes

In this subsection, we introduce two commonly used annotation schemes for anaphora resolution: MUC and MATE. There are also other schemes, for example, the Lancaster scheme [160] and the DRAMA scheme [161].

3.2.1. MUC

MUC [162,163] is one of the very first schemes, which is used for annotating the MUC [164] and the ACE [165] corpora and is still widely used these years. It is primary goal is to annotate co-reference chains in discourse, in which MUC defines and proposes to annotate the IDENTITY (IDENT) relation. Relations as such are symmetrical (i.e., if A IDENT B, then B IDENT A) and transitive (i.e., if A IDENT B and B IDENT C, then A IDENT C). Annotation is done using SGML, for example:

(8) ⟨COREF ID="100"⟩Lawson Mardon Group Ltd.⟨/COREF⟩ said ⟨COREF ID="101" TYPE="IDENT" REF="100"⟩it⟨/COREF⟩...

The annotation above construct a link between the pronoun "it" and the noun phrase "Lawson Mardon Group Ltd"..

MUC proposes to annotate co-reference chains following a paradigm analogous to anaphora resolution. Annotators are first asked to annotate markable phrases (e.g., nouns, noun phrases, and pronouns) and partition the phrases into sets of co-referring elements. This helps the annotation task achieve good inter-annotator agreement (i.e., larger than 95%).

Nevertheless, it has been pointed out by Deemter and Kibble [166] that MUC has certain flaws: MUC does not guarantee that the annotated relations are all co-referential. It includes either relation that does not follow the principle of identity-of-reference or bound anaphora. Therefore, the resulting corpus would often be a mixture of co-reference and anaphora.

3.2.2. MATE

Instead of annotating a single device INDENT, MATE [167,168] was proposed to do so-called "anaphoric annotation" which is explicitly based on the discourse model assumption [144,169–171]. The scheme was first proposed to annotate anaphora in dialogues but was then extended to relations in discourse (see [172] for more details). Such a good extensibility is a result of the fact that MATE is a *meta-scheme*: It consists of a core scheme and multiple extensions. The core scheme can be used to conduct the same annotation task as MUC and can be extended with respect to different tasks. The annotation normally uses XML, but many of its extensions use other their own formats.

3.2.3. Zero pronoun, bridging reference, and deictic reference

In addition to the "co-referential" relation discussed above, many are also interested in "hard" cases, each kind of which is often annotated as following an extension of MATE. These include the following three: (1) zero pronoun: [172] annotated (both anaphoric and non-anaphoric) ZPs in Chinese and Arabic (see Section 3.1.4); (2) bridging reference: bridging anaphora is a kind of indirect referent, where the antecedent of an anaphor is not explicitly mentioned but "associated" information is mentioned [173]. Identifying such a relation needs commonsense inference. Consider the following example from [173]:

(9) I looked into the room. The ceiling was very high.

"the room" is an antecedent of "the ceiling" because the room has a ceiling; (3) deictic reference: deixis [174] is a phrase that refers to the "speaker's position" (e.g., time, place, and situation), which is always abstracted. For example, in

(10) I went to school yesterday.

the first person pronoun "I" and the word "yesterday" are deictic references, which refer to the speaker and the day before the date when (10) was uttered, respectively. Schemes like ARRAU [175] extended MATE and is able to annotate bridging and deictic references.

[17] In linguistics, a zero pronoun is a pronoun that is implied but not explicitly expressed in a sentence.

Table 6
Anaphora resolution datasets and statistics.

Dataset	Source	#Samples	Reference
MUC	WSJ	200	Chinchor and Sundheim [164]
ACE	News	1,800	Doddington et al. [165]
GNOME	Multi-domain	505	Poesio [176]
OntoNotes	Multi-domain	4,560	Hovy et al. [177]
WSC	Manually written	285	Levesque et al. [178]
DPR	Manually written	1,880	Rahman and Ng [179]
GAP	Wikipedia	4,454	Webster et al. [180]
NP4E	Reuters	104	Hasler et al. [181]
ECB+	News	982	Cybulska and Vossen [182]
ARRAU	Multi-domain	552	Poesio and Artstein [183]

3.3. Datasets

As we discussed when we introduced annotation schemes in Section 3.2, there is no clear cut between co-reference and anaphora in computational linguistics research. We hereby review either mainstream corpora utilized in Anaphora Resolution or co-reference resolution, while being mindful of the scope of each of them. The datasets and their statistics are summarized in Table 6.

The 6th version of MUC [MUC-6, 164] is the first corpus that enables the co-reference resolution, where the task of co-reference resolution and the MUC annotation scheme was first defined. Its texts are inherited from the prevision MUCs and are English news. An example of MUC-6 is shown in List (8). Chinchor [184] updated MUC-6 in 2001 and construct the MUC-7/MET-2 corpus. MUC-7 was designed to be multi-lingual (NB: data in Chinese and Japanese are included in MET-2, which has been considered as a part of MUC-7) and to be more carefully annotated than MUC-6 by providing annotators with a clearer task definition and finer annotation guidelines.

ACE is a multi-lingual (i.e., English, Chinese, and Arabic) multi-domain co-reference resolution corpus [165]. In terms of co-reference resolution, it was built with the same purpose as MUC[18] and they same problems pointed by Deemter and Kibble [166] (see Section 3.2 for more discussion). In addition to MUC and AEC, there are works following the MUC scheme, while targeting domains other than news, which include GENIA [185], GUM [186], and PRECO [187].

The GNOME corpus was first proposed to investigate the effect of salience on language production (see Section 3.1.3 and [176,188] and then be used to develop and evaluate anaphora resolution algorithms [189,190] targeting especially the bridging reference resolution, in the course of which the MATE scheme was introduced (see Section 3.2). GNOME is an English multi-domain corpus. The initial GNOME corpus [191] consists of data from the museum domain (building on the SOLE project [192]) and patient information leaflets (building on the ICONOCLAST project), which is then expended to include tutorial dialogues [176]. GNOME followed the MATE scheme. Each noun phrase is marked by an ⟨*ne*⟩ and its anaphoric relations (marked by) are annotated separately, for example:

```
⟨ne ID="ne07" ... ⟩
Scottish-born, Canadian-based jeweller, Alison Bailey-Smith
⟨/ne⟩
...
⟨ne ID="ne08"⟩ ⟨ne ID="ne09"⟩Her⟨/ne⟩ materials⟨/ne⟩

⟨ante current="ne09"⟩
⟨anchor ID="ne07" rel="ident" ... ⟩
⟨/ante⟩
```

OntoNotes [177] is a multi-lingual (i.e., English, Chinese, and Arabic) multi-domain dataset. It is one of the most commonly used anaphora/co-reference resolution and was used in the CoNLL 2012 shared task [172]. It was annotated following an adapted version of the MATE (named M/O scheme by Poesio et al. [13]. Though it has been widely used in co-reference resolution tasks, many of its relations are not co-reference. For example, bound anaphora frequently appear (see the start of this section for more discussion). Additionally, OntoNotes annotates ZPs in its Chinese and Arabic portions (see Section 3.1.4). There are other corpora following M/O, but targeting different domains, including the biomedical (e.g., CRAFT [193]), Wikipedia (e.g., GAP [180] and WikiCoref [194]), and literary text (e.g., LitBank [195]); and different anaphorical phenomena, including bridging anaphora (e.g., ISNOTE [196]), style variation (e.g., WikiCoref [194]), and ambiguity (e.g., GAP [180]).

ARRAU is an English multi-domain (i.e., dialogue, narrative, and news) anaphora resolution dataset, annotated following the MATE scheme [183,197]. However, different from other corpora that also follow MATE, ARRAU extended MATE to annotate anaphoric ambiguity explicitly (recall that MATE is a meta-scheme). Poesio and Artstein [183] introduced the *Quasi-identity* relation, which is used for the situation when co-refer is possible but not certain by annotators and allowed each anaphor to have two distinct interpretations. In the example sample below, the footnote "1,2" of the anaphor "it" means ambiguity exists and it can either refer to 'engine E2' or "the boxcar at Elmira".

(u1) M: can we .. kindly hook up ... uh ... [engine E2]$_1$ to [the boxcar at Elmira]$_2$
(u2) M: +and+ send [it]$_{1,2}$ to Corning as soon as possible please

The Winograd Scheme Challenge [WSC, 178] focuses on the "hard" cases of CR, which often require lexical and commonsense knowledge. It can be traced back to Terry Winograd's minimal pair [198]:

(11) a. **The city council** refused the demonstrators a permit because **they** feared violence.
b. The city council refused **the demonstrators** a permit because **they** advocated violence.

The antecedent of "they" changes from "the city council" to "the demonstrators" from a to b. Levesque et al. [178] introduced the WSC benchmark consisting of hundreds of such minimal pairs. Since then, many larger-scale WSC-like corpora have been constructed. This includes the DPR corpus [179], the PDP corpus [199], and the Winogrande corpus [200]. Following a similar paradigm, GAP [180], Winogender [201] and Winobias [202] were proposed for "hard" cases that link to gender bias.

NP4E [181] and ECB+ [182] are corpora for investigating cross-document co-reference. They annotated both entities and events co-reference and both within and cross-document co-reference. These corpora were built by starting from a set of clusters of documents, the documents of each of which describe the same fundamental events.

The corpora mentioned above are all in English, some of which have Chinese and Arabic portions. There are anaphora/co-reference resolution corpora that focus on languages other than them. These include ANCOR [in French, 203], ANCORA [in Catalan and Spanish 204], COREA [in Dutch 205], NAIST [in Japanese 206], PCC [in Polish 207], PCEDT [in Czech 208], and TUBA-DZ [in German 209].

3.4. Knowledge bases

Both lexical and world knowledge are useful for anaphor interpretation. See the following examples from [210]:

(12) a. There was a lot of **Tour de France riders** staying at our hotel. Several of **the athletes** even ate in the hotel restaurant.

[18] Though, in terms of entity recognition, they don't have the same purpose.

Table 7
Useful knowledge bases for anaphora resolution.

Name	Knowledge	#Entities	Structure
WordNet	Lexical	155,327	Graph
COW	Lexical	157,112	Graph
ODW	Lexical	92,295	Graph
AWN	Lexical	≈10,000	Graph
Wikipedia	World	13,489,694	Unstructured
Wikidata	World	100,905,254	Graph
DBpedia	World	≈4,580,000	Graph
Freebase	World	≈2.4 B	Graph
YAGO	World	4,595,906	Graph
WikiNet	World	3,347,712	Graph
OMCS	World	62,730	Graph
Medical-KG	World	22,234	Graph

b. She was staying at **the Ritz**, but even that **hotel** didn't offer dog walking service.

We need the lexical knowledge that indicates "riders" are "athletes" while need the world knowledge of the fact that "Ritz" is a "hotel" (see Table 7).

WordNet provides lexical knowledge of English [211], including lexical entries (e.g., meaning, part-of-speech, etc.) and relations (e.g., synonyms, hyponyms, and meronyms, etc.) among them (see Table 5).

Wikipedia has been an important world knowledge source for many anaphora/co-reference resolution systems. These knowledge bases consist of documents from Wikipedia as well as related meta-data. Typical examples include bases from those directly dumped from raw Wikipedia documents[19] to better-structured ones, such as Wikidata [212], DBpedia [213], and Freebase [214].

Knowledge Graphs have become popular in anaphora/co-reference resolution tasks because bases that build on raw Wikipedia are needed to be further processed (e.g., entity and relation extraction) before use. Popular knowledge graphs include those that build on Wikipedia (e.g., YAGO [215] and WikiNet [216]), that are about Commonsense (e.g., OMCS [217]), and that are about expert knowledge (e.g., Medical-KG [218]).

Search Engines, e.g., Bing and Google were also used by a few works (e.g., [219]) to "hunt" knowledge for the target entities in order to resolve hard anaphora like those in WSC (see Section 3.3), in addition to the above knowledge bases in the strict sense.

3.5. Evaluation metrics

Vanilla Precision, Recall and F1. A plausible way to assess anaphora resolution systems is by viewing both mention detection and mention linking tasks as simple classification tasks and measuring the performance using vanilla precision, recall, and F1 scores. A good evaluation metric needs to be both interpretable and discriminative. However, unfortunately, these measures cannot meet any of these criteria [220], especially for the mention linking task as they overlook the structure of these relations (most of which are chain-structured).

MUC and Beyond. Along with MUC-6 (see Section 3.3), [221] proposed the MUC score. It computes the recall and precision of anaphora/co-reference resolution outputs by considering co-reference chains in a document as a graph. Vilain et al. [221] first defined two sets: a set of key entities $\mathcal{K}$, in which there are gold standard reference chains (NB: a chain is sometimes named as a class or a cluster), and a set of response entities $\mathcal{R}$, in which there are system generated chained. MUC score computes the recall based on the number of missing links in $\mathcal{R}$ compared to $\mathcal{K}$, formally:

$$\text{Recall} = \frac{\sum_{k_i \in \mathcal{K}} \left(|k_i| - |p(k_i, \mathcal{R})|\right)}{\sum_{k_i \in \mathcal{K}} \left(|k_i| - 1\right)} \tag{4}$$

where $|k_i|$ is the number of mentions in the chain k_i and $p(k_i, \mathcal{R})$ is the set of partitions that is constructed by intersecting k_i with $\mathcal{R}$. The computation of MUC precision is done by switching $\mathcal{K}$ and $\mathcal{R}$. However, it has been pointed out that MUC has certain flaws: on the one hand, since MUC is merely building on mismatches of links between the two sets, it is not discriminative enough [222,223]. For example, it does not tell the difference between an extra link between two singletons or two prominent entities. On the other hand, [223,224] argued that MUC prefers singletons. For instance, if we merge all mentions in OntoNotes into singletons, the resulting MUC will be higher than that of the SOTA [220].

Many metrics beyond MUC have been proposed by measuring recall and precision using mentions instead of links. Bagga and Baldwin [222] proposed B^3, which considers the fractions of the correctly identified mentions in $\mathcal{R}$:

$$\text{Recall} = \frac{\sum_{k_i \in \mathcal{K}} \sum_{r_j \in \mathcal{R}} \frac{|k_i \cap r_j|^2}{|k_i|}}{\sum_{k_i \in \mathcal{K}} |k_i|} \tag{5}$$

The precision is also computed by switching $\mathcal{K}$ and $\mathcal{R}$. As pointed by Luo [223] and Luo and Pradhan [225], B^3 still cannot fully properly handle singletons and, additionally, repeated mentions. To solve this, [223] proposed CEAF to incorporate measures of similarities between entities:

$$\text{Recall} = \frac{\sum_{k_i \in \mathcal{K}^*} \phi(k_i, g(k_i))}{\sum_{k_i \in \mathcal{K}} \phi(k_i, k_i)} \tag{6}$$

where $\mathcal{K}^*$ is the set of key entities that have the optimal mapping with $\mathcal{R}$, which is found by the Kuhn–Munkres algorithm, and $\phi(\cdot)$ is a similarity measure. Nevertheless, CEAF has two shortcomings: it overlooks all unaligned response entities [226] and weights entities equally [227].

In addition to above mentioned based metrics, to handle singletons, [228] proposed BLANC to also consider non-coreference/non-anaphoric links. It measures the fiction of both correctly identified co-reference links and non-coreference entities, and averages them to obtain the final score.

Moosavi and Strube [220] conducted controlled experiments and proved that all the aforementioned computations of precision and recall are neither interpretable nor reliable as they suffer from the so-called *mention identification effect*. They proposed the LEA metric, which was claimed to be able to solve the above issues from two perspectives: (1) it considers both links and mentions; (2) it weights entities with respect to their importance.

3.6. Annotation tools

Text Editors. In the early years, anaphora/co-reference were annotated using text editors or manipulation tools. For example, MUC-6 and ACE were annotated using plain text editors while GNOME was annotated using the XML manipulation tool developed by the University of Edinburgh.[20]

Co-reference Annotation Tools. Later, linguists and computer scientists developed software that enables multi-layer annotation. The software that is designed for annotating co-reference or allows the annotations of relations between phrases can be used for anaphora/co-reference annotation tasks. For example, ARRAU and PCC used MMAX2, which is a free, extensible, general-purpose, and desktop-based annotation tool. It allows users to annotate relations using fields in a form, and the form is customizable. The NP4E project used PALinkA and ECB+ used CAT [229]. Both of them were designed for the event and reference annotation. More recently, co-reference annotation tools that provide better visualization, allow drag-and-drop

[19] https://dumps.wikimedia.org/

[20] http://www.ltg.ed.ac.uk/software/

annotation, and offer post-annotation analysis have been built. Typical examples include CorefAnnotator [230], which is open-sourced and desktop-based, SCAR [231], which is open-sourced and web-based, and LightTag, which is not fully free but provides good online teamwork services.

Annotation Tools with Advanced Functionalities. Some annotation tools provide extra services that help to make sure the annotation procedure is fast and reliable. We classify these services into three categories: (1) External Knowledge: BRAT [232] and INCEpTION [233] integrate external knowledge bases, e.g., Freebase and Wikidata (see Section 3.4). Once an annotator identifies an entity, these tools would search the linked base and return related entry; (2) Pre-trained Models: Tools such as TagEditor, Togtag, INCEpTION, and MyMiner [234] can call embedded pre-trained entity recognition models so that they can suggest positions of possible name entities during annotation, in which MyMiner was designed specifically for the medical domain (see [235] for an overview of annotation tools for medical NLP). Additionally, beyond name entities, TagEditor and INCEpTION can also suggest potential reference chains based on their integrated pre-trained co-reference resolvers, enabling active learning for anaphora/co-reference resolution; (3) Cross-document Annotation: using CROMER [236] and CoRefi [237], annotators can tag, link, or update entities across multiple documents. This is done by allowing annotators to cluster and annotate documents based on topics.

3.7. Methods

3.7.1. Rule-based methods

A. Linguistically-inspired Approaches

Like many other tasks in NLP, early works on anaphora resolution built on rules that are rooted cognitively and linguistically. Here, the term “early” represents the age when systematic evaluations of anaphora resolution, e.g., MUC, had not been introduced. The very first algorithm is the naive algorithm proposed by Hobbs [238]. It first does a breadth-first search from the parse tree of the sentence to search for identifying mentions and links mentions based on constraints introduced in Section 3.1.1. Later on, a series of anaphora resolution systems were proposed together with computational investigations of the effect of salience (see Section 3.1.3). Based on a set of factors that proved to influence salience, [239] introduced rules that are used to compute the expected focus of discourse and rules that are used to interpret anaphora. As a matter of fact, this work was built on the “centering view” rooted from [240], which suggests that, during anaphora resolution, the searching of antecedents should be restricted to the set of centered entities. It could be seen as a prototype of the idea of “center of salience” of the centering theory (see Section 3.1.2), but the rules proposed by Sidner [239] are extremely complex.

Starting from [239], [241] focused on the rules about salience and developed a system coined Shallow Processing Anaphor Resolver (SPAR). SPAR maintains linguistically-inspired rules as domain knowledge and does commonsense inference over them. As pointed out by Carter [241], since maintaining domain knowledge and reasoning rules is expensive, SPAR made them as simple as possible. That is why it was called “shallow processing”. Carter assessed SPAR on a set of 322 test samples and found that SPAR could successfully resolve 93% pronominal anaphors and 87% non-pronominal anaphora. Hobbs et al. [242] formalized commonsense inference in anaphora resolution as abduction and introduced TACITUS. To do abduction, in TACITUS, knowledge (i.e., rules) is maintained in formal logic (first-order predicate logic in this case). Focusing on salience, [243] proposed the Resolution of Anaphora Procedure (RAP) algorithm. After selecting a set of candidate antecedents based on semantic and syntactic constraints, RAP contains a rule-based procedure for assigning values to several salience parameters, which are then used for resolute anaphors. An assessment on 360 hand-crafted texts containing pronouns showed RAP defeated the naive algorithm by 2%.

Also starting from [239], there were subsequent works that extended the idea of “focus” on the basis of the introduction of the concept of “centering”. Brennan et al. [244] introduced the BFP algorithm for anaphora resolution, which roughly has three stages: (1) construct a set of candidate antecedents with accordance to the rules of the semantic constraint; (2) filter and classify the candidates based on which action a candidate belongs to in centering theory (see Section 3.1.2); and (3) select the best candidate in according to a predefined preference over the actions. One limitation of the BFP algorithm is that its final choice is merely based on a linear preference order. To optimize this selection process, [245] marries BFP with the optimality theory. Another limitation is that, by only considering the center theory, BFP overlooked a key pattern of how human resolute pronouns, namely, incremental resolution [140]. In response to this problem, [246] proposed the Left-to-Right Centering (LRC) algorithm, which is an incremental resolution algorithm that adheres to centering constraints. An evaluation on the New York Time corpus [247] suggests that LRC outperformed both BFP and the naive algorithm.

B. Knowledge-poor Approaches

After the introduction of the MUC-6 shared task, anaphora resolution systems are able to be evaluated on a large scale. However, the trade-off is that the anaphora resolution systems can no longer access inputs that are annotated with gold-standard semantic and syntactic knowledge. Building on this setting, “knowledge-poor” approaches were proposed and most systems of this kind prefer rules that have high precision but do not rely on knowledge. The most influential work is CogNIAC [248], which is a heuristic precision-first anaphora resolver that relies on rules that are almost always true. For example, CogNIAC contains a rule saying *if there is just one possible antecedent in entire the prior discourse, then that entity is the antecedent*. Its rules were selected based on the precision tested on a set of test sentences. It is worth noting that rules in CogNIAC are still used in many SOTA practical anaphora resolution systems (e.g., the Stanford Deterministic Coreference Resolver [249]).

C. Approaches with Approximate Knowledge

As pointed out by Poesio et al. [13], this encourages two major changes in anaphora resolution: one this that instead of relying on perfect knowledge and doing reasoning on it, anaphora resolution systems started to syntactic parsers and approximate knowledge like WordNet. The other is that the focus of anaphora resolution models moved from being aware of only pronouns to all kinds of nominal phrases (that function as referring).

Kameyama [250] proposed to resolve anaphors that are proper names, descriptions, and pronouns. It relies on syntactic and semantic constraints, but the related information came from a syntactic parser and morphological filter based on person, number, and gender features. Later on, approaches that marry rules with WordNet were introduced [251,252]. They made use of heuristic rules (as in CogNIAC), some of which consider lexical information from WordNet.

The most famous rule-based anaphora resolution system is the one proposed by Haghighi and Klein [253], which is still frequently used as a strong baseline in today’s research on anaphora resolution. In addition to aforesaid syntactic and semantic constraints, [253] makes full use of the parse trees. For example, it contains rules that rely on the distance between mentions, which is obtained from computing the shortest path between two mentions in the parse tree. It also uses Wikipedia as a resource for acquiring semantic knowledge of each entity.

One limitation of heuristic-based systems is that lower precision features often overwhelm higher precision features. In response to this, more recently rule-based systems [249,254] categorized rules into sieves and made decisions with an ordered set of rules. These works are often called multi-sieve approaches.

3.7.2. Statistical-based methods

The introduction of large-scale benchmarks also encourages the trend of using machine learning techniques in anaphora resolution. Basically, these learning-based models treat anaphora resolution as a series of classification problems. We categorize them on the basis of how they define the classification task.

A. Mention-pair Models

Mention-pair models train a classifier to determine whether two mentions co-refer or not. It was first introduced by Aone and Bennett [255] and then perfected by Soon et al. [256]. To build a mention-pair model, there are five steps:

1. Identifying Mentions: As a practical anaphora resolution model, the first step of this framework is to identify mentions. Soon et al. [256] break down the mention identification into two stages: they first used three statistical sequence taggers (which is a Hidden Markov Model [257]) to do part-of-speech tagging, noun phrase identification, and name entity recognition, respectively. The outputs of them are noun phrases as well as name entities. Then, they designed rules to recognize nested noun phrases based on the identified noun phrases. For each discourse, the resulting set of mentions is the union of noun phrases, name entities, and nested noun phrases. In later works, this module was replaced by more advanced sequence taggers, e.g., conditional random field. See [258] for a survey.
2. Feature Engineering: Akin to many statistical models, feature engineering is always needed. Soon et al. [256] made use of not only syntactic and semantic features as usual but also lexical features with the help of WordNet. In addition to [256], many works used knowledge bases for feature engineering (e.g., [259,260]). In 2008, [261] found that a simple model with good feature engineering can defect the SOTA model at that moment.
3. Generating Training Examples: They used a heuristic-based method to generate training pairs (i.e., a pair of positive and negative examples). More specifically, a positive instance consists of an anaphor A_1 and its closest preceding antecedent A_2 while a negative instance consists of the same anaphor A_1 and the mention that intervenes A_1 and A_2. There has been a number of modifications to this strategy. For example, [262] forced that A_1 can only be a non-pronominal once A_2 is also a non-pronominal. Harabagiu et al. [263]; Ng and Cardie [264]; Strube et al. [265]; Yang et al. [266] further enhanced this process by applying rule-based or learning-based filters.
4. Building a Classifier: In this step, statistical machine learning techniques have been used. These include decision trees [256,267], random forests [268], Max Entropy classifier [247, 269], and memory-based learning [270].
5. Generating Co-reference Chains: The last step is to partition these anaphora into co-reference chains. Normally, clustering techniques are used in this step. These include closest-first clustering [256], best-first clustering [262], correlational clustering [271], and graph partitioning algorithms [272,273].

B. Entity-Mention Models

As a matter of fact, the task mention-pair anaphora resolution is counter-intuitive from the perspective of linguists and cognitive scientists. Additionally, [13] pointed out that mention-pair models also overlook features of entities [274]. In response to this, entity-mention models were proposed. They directly link mentions to entities by clustering. Specifically, [275] trained a model to classify whether a mention belongs to a partially constructed cluster. However, according to the evaluation by Luo [223], the performance of the models of this kind is not comparable to mention-pair models.

C. Mention-Ranking Models

Another problem of mention-pair models is that they only do binary classification without comparing different potential antecedents. To remedy this, [276] proposed an entity-ranking model, replacing the binary classification loss with a ranking loss. Rahman and Ng [277] combined entity-ranking strategy with the entity-mention model, yielding SOTA performance at that moment.

3.7.3. Neural anaphora resolution

A. Conventional Deep Learning Models

Wiseman et al. [278] was the first to use deep neural networks in anaphora resolution. It is a non-linear mention-ranking model. Instead of conjunction features (as in statistical models), the model of Wiseman et al. [278] uses a neural network to learn feature representations as an extension to the mention-ranking model. They defined two feature vectors, each of which is obtained from pre-training the model on any of the sub-tasks of anaphora resolution, namely, mention identification and mention linking. The final decision is made through a non-linear classification, based on these features. Both [279] and [280] augmented the work of Wiseman et al. [278] by inducing global features, but they followed different schemes. Wiseman et al. [279] ran a recurrent neural network (RNN) to encode the representation of each sequence of mentions corresponding to an entity (i.e., a cluster) in the history. Whereas, [280] first used a feed-forward neural network to encode each mention-pair of an entity and computed the entity representation by pooling over all mention-pairs. Later on, [281] extended their previous work [282], which built up co-reference chains with agglomerative clustering. Each mention starts in its own cluster and then pairs of clusters are merged using imitation learning (a type of reinforcement learning technique) by assuming merging clusters are actions. Clark and Manning [281] replaced imitation learning with deep reinforcement learning. Liu et al. [283] proposed a multi-task learning framework for mention detection and mention linking tasks, because they found that the learning of mention detection task can enhance the learning of dependent information of input tokens, which is complimentary for mention linking detection. Such an approach achieved comparable performance to [284] with only 0.05% WIKICREM training samples.

B. End-to-End Models

A significant benefit of employing deep learning models lies in their capacity to operate without the requirement of handcrafted features, thus enabling the creation of end-to-end (End2End) systems. Lee et al. [285] proposed the first End2End anaphora resolution system. It needs no human-craft feature or parser and, more importantly, it learns to process mention identification and linking tasks jointly. To this end, the fundamental idea is to first view all spans in the previous discourse as candidate antecedents and do mention ranking (NB: it was called span ranking in [285] as the spans it sent for rank are not always mentions). The inputs pass through an RNN and each span is represented by the concatenation of the RNN hidden states of the first token and the last token as well as the weighted sum of all tokens in the span using the attention mechanism [286]. The final decision of each pair is made using a feed-forward neural network. One limitation of this method is that since it searches over all possible spans, the search space would be extremely large. To remedy this, candidate spans are pruned by limiting the maximum span width, the number of spans per word, the maximum number of antecedents, and the length of input documents. This End2End model was tested on the OntoNotes dataset and outperformed all previous works.

Akin to mention-pair anaphora resolution systems, End2End anaphora resolution is problematic because it ranks every span-anaphor pair separately. In response to this problem, [287] introduced a higher-order coarse-to-fine inference strategy for End2End anaphora resolution models (henceforth, C2F-AR), which, in short, does cluster ranking. It infers in an iterative manner. The antecedent distributions are used to update the span representations before doing inference, enabling later decisions conditioned on previous decisions. C2F-AR uses a coarse factor that can further prune candidate span during this higher-order inference, More recent works focused on either improving span representations or selecting candidate spans. For example, [288] used a two-layer bi-directional RNN and combined the representations of adjacent sentences in order to improve span representation with cross-sentence dependency information. Zhang et al. [289] proposed to enrich the span representations by training a mention identification model jointly assigning each candidate span an antecedent score. For each pair of spans, [290] replaced span representations with a combination of lightweight bilinear functions between pairs of endpoint token representations. Wu et al. [291] formalized the End2End anaphora resolution as a question-answering task. A query is produced for each entity and predicts the positions of all spans in the co-reference chain.

C. Knowledge-based Models

Analog to classical rule-based and statistical-based approaches, works on neural anaphora resolution models also seek to integrate knowledge. In terms of the use of open knowledge bases, [292] used world knowledge to compute rewards for reinforcement learning-based anaphora resolution models. More specifically, they submitted the predictions to an OpenIE system and compared the predicted anaphora with the knowledge to compute the reward. Zhang et al. [293] extracted knowledge triples related to each entity from knowledge graphs and used them to enrich span representations using a knowledge attention module.

It has been pointed out that pre-trained language models are knowledge bases [294]. Many recent anaphora resolution models have incorporated pre-trained language models, including BERT [8], SpanBERT [295], and CorefBERT [296].

There has been a line of work focusing on addressing mention linking in WSC-like corpora (see Section 3.3). As aforementioned, resolving these “hard” cases needs reasoning with world knowledge. Works of this line incorporate either external knowledge bases [219] or pre-trained language models [284,297].

3.7.4. Anaphoric zero pronoun resolution

As mentioned in Section 3.1.2, “cool” languages (e.g., Chinese, Japanese, Korean, and Arabic) contain anaphoric zero pronouns (AZPs), and many works have focused on resolving AZPs. As with other anaphora resolution tasks, early works on AZP resolution (AZPR) used rule-based approaches and statistical approaches. Theoretically, these works are built on the fact that speakers process zero pronouns (ZPs) in the same way as pronouns [298]. Early on, most of the works are for Japanese because of the NAIST corpus [206], in which AZPs are annotated. Kameyama [299]; Okumura and Tamura [300] used center theory-based approaches for AZPR in Japanese. Statistical-based approaches were proposed with a focus on exploring useful features, including syntactic pattern features [301], heuristic rules [302], and features that had been considered in anaphora resolution systems [303–308]. Meanwhile, there were also a number of Korean AZPR systems building on the Korean portion of Penn Treebank [309,310].

Later on, the development of systems for Chinese [159,311–314] and Arabic AZPs became active after the introduction of OntoNotes [315]. From [316], AZPR systems also went into the age of deep learning. Most of the works were for Chinese AZPR, including approaches that use deep feedforward neural networks [316], RNNs [317,318], attention network [319], memory network [320], deep reinforcement learning [321] and BERT [322].

The training of AZPR systems shares the problem of lacking annotated training data. For example, the AZPR largest corpus, i.e., the Chinese portion of OntoNotes, contains only 12,111 AZPs. To incorporate more data into training, there have been three paradigms: (1) Joint modeling: [323] and [324] proposed to train a model that resolves either AZPs and non-zero pronouns jointly; (2) Multi-linguality: [325] and [315] trained multi-lingual AZPR systems which were trained on AZPR data in multiple languages; (3) Data augmentation: [326] made use of large-scale reading comprehension dataset in Chinese to generate pseudo training data for Chinese AZPR. Aloraini and Poesio [327] augmented Arabic AZPR data by a number of augmentation strategies, e.g., back translation, masking candidate mentions, etc.

3.8. Downstream applications

3.8.1. Machine translation

Stojanovski and Fraser [328] provided the following example to illustrate how oracle anaphora singles can help machine translation systems.

(13) a. Let me summarize the novel for you.
b. It presents a problem.
c. er!@#$XPRONOUN It presents a problem.
d. Er prasentiert ein Problem.

Given the context (a) and the course sentence (b), based on the oracle anaphora information, [328] pre-pend the input sentence of machine translation with pronoun translation as shown in (c) and ask the system to translation with a target (d) in German. In this case, the pronoun “it” which refers to “novel” (in German “Roman”) is translated to “er” (the German masculine pronoun agreeing with “Roman”). Without this information, they argued that machine translation will be hard to produce “er”. The experiment on a number of Neural machine translation models suggested that would improve the BLEU scores by 4–5 points. This argumentation was strengthened by the experiments conducted by Saunders et al. [329], who concluded that NMT does not translate gender co-reference. Despite these theoretical studies, many works [330–332] focused on improving machine translation with anaphora resolution outputs. The solution is often using anaphora resolution outcomes to obtain features of each pronoun (including, gender, number, and animacy) in order to enhance the pronoun translation performance. Beyond these works, [333] proposed to use clustering scores which are used for generating co-reference chains in anaphora resolution (see Section 3.7.2) as features for re-ranking machine translation results.

There has been a long tradition of studying the impact of AZPs on machine translation systems, especially when translating from a pro-drop language to a non-pro-drop language. For example, the Japanese–English machine translation in the 1990s had already been deployed an AZPR systems [334]. Later systems followed a slightly different strategy. Instead of doing a full anaphora resolution, these systems only detect AZPs in the source language and directly translate them into the target language without further resolute them [335,336].

3.8.2. Summarization

There are two major uses of anaphora resolution in text summarization [337]. One is to help with finding the important terms while the other is to help with evaluating the coherence of the summarization. Many works have demonstrated that incorporating the information of co-reference chains contributes to both the faithfulness and the coverage of summarization systems [338–341]. Nevertheless, it is also worth noting that there are also some studies that showed that anaphora resolution had negative effects [342,343]. One possible explanation is that the effect highly depends on the task the summarization system is addressing and the performance of the anaphora resolution systems (NB: these studies have been 15 years old).

3.8.3. Textual entailment

For textual entailment, to understand the impact of anaphora resolution, [344] manually analyzed 120 samples in the RTE-5 development set [345]. They found that for 44% samples anaphora relations are mandatory for inference and for 28% sample anaphora optionally support the inference. Based on this fact, many systems that got involved in the RTE challenge made use of anaphora resolution. Nevertheless, since anaphora resolution systems at that moment were not strong enough, errors they made would propagate to downstream textual entailment systems [346,347]. As a consequence, the contribution of anaphora resolution was negative or not significant [348,349].

3.8.4. Sentiment computing

For sentiment computing, [133] listed two situations when anaphora resolution can help. One is when doing sentiment analysis on online reviews, a characteristic of them is that online reviews often focus on a particular entity and, therefore, the mentions often in less elaborated forms (e.g., pronouns). Resolution of these mentions can chain them into a global entity and, hence, improve the sentiment analysis performance. The other is that anaphora resolution can also be used in fine-grained aspect-based sentiment analysis. Anaphora resolution plays a pivotal role in this task by facilitating the clustering of entities into distinct aspects. This, in turn, aids in the extraction of sentiments and opinions associated with each aspect.

The contribution of anaphora resolution in sentiment computing tasks can be summarized as follows: it enables discourse-level sentiment analysis by linking mentions from different sentences. Many efforts have been carried out to demonstrate such an ability for anaphora resolution. Nicolov et al. [350] conducted systematic experiments to understand the impacts of anaphora resolution on sentiment analysis. Specifically, they tried to incorporate anaphora information into a number of sentiment analysis models and assessed them on varieties of datasets. They concluded that, on average, anaphora resolution can boost sentiment analysis performance by 10%. Based on this finding, sentiment analysis systems that are assembled with anaphora resolution have been proposed [351–353].

3.9. Summary

Anaphora resolution has been explored extensively by theoretical linguists, psycholinguists as well as computational linguistics. It is the manifest of structural semantics because the meaning of an anaphor elucidates the syntactic relationship between the anaphor and its antecedent. Early anaphora resolution models were inspired by theories and findings in linguistics, such as the theory of syntactic and semantic constraints from theoretical linguistics and the findings about factors that influence the choice of referential form from psycholinguists. Later on, by marrying these theories with computational models, linguists also gained insights regarding the comprehension and production of anaphora from anaphora resolution systems. For instance, we could understand better how each salience factor contributes to the use of anaphora through the importance analysis of a computational model that considers the factor. Most recently, though most computational works focus on building End2End anaphora resolution systems based on deep learning techniques, linguistic theories about anaphora are still proven to play vital roles [354]. Dataset is core for either practical or theoretical anaphora resolution research. Though many annotation schemes and datasets have been introduced, we found that they share two limitations: one is that due to the fact that anaphora is a complex concept, annotations of anaphora resolution datasets are always imperfect [166]. The other is the lack of wide-coverage datasets that covers all kinds of anaphora. Finally, we found that anaphora resolution is useful in many downstream tasks, including major tasks of both natural language understanding and natural language generation. It is always utilized as a producer of additional features for downstream tasks. Different from other tasks in this survey, we rarely see how anaphora resolution techniques help boost the explainability of downstream models, apart from the work of Saunders et al. [329]. We also have not observed that anaphora resolution techniques are used for constructing datasets for downstream tasks.

3.9.1. Technical trends

As seen in Table 8, there are two clear technical trends. One is that the research interest in the realm of anaphora resolution has shifted from machine learning-based or rule-based anaphora resolution to neural approaches, especially the End2End neural anaphora resolution, which does mention identification and linking simultaneously. Another one is that, as previously elucidated in Section 3.7, there exist distinct shortcomings associated with each of the task formulations such as mention pair, entity mention, and mention ranking. Consequently, a recent tendency is to employ higher-order inferences [287] to directly rank clusters or entities, which allows for the incorporation of benefits from all the formulations. To sum up, the SOTA anaphora resolution models are often *End2End cluster ranking models*.

Most recent advances tended to further improve this paradigm from two angles, namely reducing the search space as an End2End anaphora resolution searches across all possible spans in its inputs for antecedents [291]; and equipping anaphora resolution systems with knowledge (which, recently, often large-scale pre-trained language models) to boost their ability of reasoning [295,355]. Furthermore, recent investigations on anaphora resolution have also led to advancements in various deep learning paradigms. Deep reinforcement learning and multi-task learning were employed for obviating the need for language-orientated hyperparameter tuning [281], investigating the enduring impact of pronoun-candidate antecedent pairs [321], and enhancing the dependency learning of mention pairs [283].

Meanwhile, there were also certain efforts that concentrated on resolving "hard" cases and multi-linguality in anaphora resolution. As for the former one, people were aware of the models' capacity to resolve ambiguous pronouns and biases (especially, gender bias) learned by anaphora resolution models [178,201]. The SOTA models of this line of work are often assembled with knowledge bases [219] or pre-trained language models [284]. As for the latter one, multi-lingual anaphora resolution systems were developed in order to either, theoretically, unify the theory of reference for different languages [356], or, practically, enrich the datasets for low-resource anaphora resolution languages or tasks (e.g., AZPR; [315]).

In addition to these two trends for developing practical anaphora resolution systems, there is also a long tradition of studying how human beings understand and use anaphors with the algorithms introduced in this section from the age of rule-based methods [239,241] to the most recent deep learning based methods [354,357].

3.9.2. Application trends

Many demonstrations were carried out approximately 15 years ago to validate the necessity of anaphora resolution for both language generation and understanding downstream tasks [337,344,350,358,359]. Nevertheless, practically, at that moment, anaphora resolution often had negative effects [342,343,348,349]. This is mainly because anaphora resolution systems were not powerful enough and errors they made may propagate to their downstream tasks.

Recently, with significant advancements in the capabilities of anaphora resolution systems, more and more anaphora resolution systems have been used for providing anaphora information for downstream tasks (see Table 9). In short, anaphora resolution helps its downstream applications mainly in two ways. It links noun phrases in different sentences. As a consequence, these applications have better performance in comprehending discourse-level information. On the other hand, linking noun phrases helps downstream applications to do higher-level reasoning, e.g., extracting global entities [133] and recovering the ellipses [360]. Most downstream task models utilize anaphora resolution as an additional feature to improve task performance. However, we did not see how anaphora resolution techniques help to explain how and why anaphora is used in a certain context.

3.9.3. Future works

Developing robust annotation schemes. Current annotation schemes for anaphora practically work fine but are theoretically problematic as there is no unified rule of what is remarkable, and no clear cut between co-reference and anaphora (though there is a clear boundary between them in linguistic theory). Annotation schemes so far are imperfect to improve the practicality so that large anaphora/co-reference resolution datasets (that can be used for training and assessing data-driven anaphora resolution systems) could be constructed.

Table 8
A summary of representative anaphora resolution techniques. Note that [277] reported that the performance of [276] was 57.7% CEAF-F and that CoNLL-F is the average of MUC, B3, and CEAF scores. Stat. denotes statistics. DL denotes deep learning. AZPR denotes Anaphoric Zero Pronoun Resolution. Cha. Emb. denotes character embedding. ACE denotes automatic content extraction. MTL denotes multi-task learning. DRL denotes deep reinforcement learning.

Task	Reference	Feature	Framework	Dataset	Score	Metric
Rule-based	Carter [241]	Salience	Logic rules	Self-collected dataset	93.00%	Acc
	Lappin and Leass [243]	Salience	Logic rules	Self-collected dataset	85.00%	Acc
	Brennan et al. [244]	Semantic constraints	Centering theory	New York Times	59.40%	Acc
	Tetreault [246]	Semantic constraints	Centering theory	New York Times	80.40%	Acc
	Baldwin [248]	Syntactic, Semantic, Discourse	Logic rules	Self-collected dataset	77.90%	Acc
	Liang and Wu [252]	WordNet	Logic rules	Brown Corpus	77.00%	Acc
	Haghighi and Klein [253]	Syntactic, Semantic	Logic rules	ACE	79.60%	MUC-F
Stat.-based	Soon et al. [256]	Syntactic, Semantic, WordNet	Mention-pair	MUC-6	62.60%	MUC-F
	Cardie and Wagstaff [275]	Lexical, Syntactic, Semantic	Entity-Mention	MUC-6	64.90%	MUC-F
	Denis and Baldridge [276]	Linguistic & Positional	Mention-ranking	ACE	67.00%	CEAF-F
	Rahman and Ng [277]	Lexical, Syntactic, Semantic	Mention-ranking	ACE	60.80%	CEAF-F
DL-based	Wiseman et al. [278]	Syntactic, Semantic	Mention-rank., DNN	OntoNotes	82.86%	Acc
	Wiseman et al. [279]	Syntactic, Semantic, Global feature	Mention-rank., RNN	OntoNotes	64.21%	CoNLL-F
	Clark and Manning [281]	Syntactic, Semantic	DRL	OntoNotes	65.73%	CoNLL-F
	Clark and Manning [280]	Syntactic, Semantic, Global feature	Mention-ranking, DNN	OntoNotes	65.52%	CoNLL-F
	Lee et al. [285]	Word & Cha. Emb.	End2End, LSTM, DNN	OntoNotes	68.80%	CoNLL-F
	Lee et al. [287]	ELMo	End2End, LSTM, DNN	OntoNotes	73.00%	CoNLL-F
	Zhang et al. [289]	Glove & Cha. Emb.	BiLSTM, Joint learning	OntoNotes	69.20%	CoNLL-F
	Joshi et al. [355]	BERT	Lee et al. [287]	OntoNotes	76.90%	CoNLL-F
	Joshi et al. [295]	SpanBERT	Lee et al. [287]	OntoNotes	79.60%	CoNLL-F
	Wu et al. [291]	SpanBERT	QA	OntoNotes	83.10%	CoNLL-F
	Kocijan et al. [284]	BERT_WikiCREM	DNN	DPR	84.80%	Acc
	Liu et al. [283]	BERT	Transformer, MTL	DPR	84.58%	Acc
AZPR	Okumura and Tamura [300]	Salience	Center Theory	Self-collected dataset	78.30%	Acc
	Sasano et al. [307]	Salience	Probalistic	Self-collected dataset	39.10%	F1
	Chen and Ng [316]	Syntactic, Lexical	DNN	OntoNotes	52.20%	F1
	Yin et al. [317]	Word2Vec, Global	RNN	OntoNotes	53.60%	F1
	Yin et al. [321]	Word embedding	DRL	OntoNotes	57.20%	F1
	Song et al. [322]	BERT	DNN, MTL	OntoNotes	58.49%	F1

Table 9
A summary of the representative applications of anaphora resolution in downstream tasks. ✓ denotes the role of anaphora resolution in a downstream task.

Reference	Downstream task	Feature	Explain.
Le Nagard and Koehn [330]	Machine translation	✓	
Hardmeier and Federico [331]	Machine translation	✓	
Miculicich and Popescu-Belis [333]	Machine translation	✓	
Saunders et al. [329]	Machine translation	✓	✓
Steinberger et al. [337]	Summarization evaluation	✓	
Bergler et al. [338]	Summarization	✓	
Liu et al. [341]	Summarization	✓	
Agichtein et al. [347]	Textual entailment	✓	
Jakob and Gurevych [351]	Sentiment computing	✓	
Ding and Liu [352]	Sentiment computing	✓	

In exchange, the resulting corpora are imperfect in terms of both quality (i.e., some annotated relations might not be anaphoras) and coverage (i.e., some kinds of anaphora are not covered). On a different note, anaphora resolution, which can also be seen as a pragmatics task, disagreement on how an anaphora is interpreted happens across different readers [361]. Nonetheless, many datasets resolve disagreements through majority voting, while only a few works explicitly annotated ambiguities, which are the causes of the disagreements (e.g., [183]). In aggregate, it is plausible to design a scheme (probably by extending MATE) that not only handles disagreements but also balances quality, practicality, and coverage. Furthermore, it is important to empirically investigate how the errors and limitations inherent in the annotation scheme can impact the performance of anaphora resolution systems.

Anaphora resolution evaluation. Analogue to the disagreements in the anaphora annotation, one can expect that, for a single mismatch between an output and a reference answer, it might be an error for some readers but not an error for the rest. For different mismatches, they might have different severity. The impact of severity of errors has been studied for the production of reference (see [362]; e.g., saying “a woman is a man” is more serious than saying “a red coat is pink”), but it has never been explored in the realm of anaphora resolution. This said, roughly computing the overlaps between model outputs and reference outputs might be problematic. On the one hand, due to discrepancies and varying degrees of errors in anaphora resolution, human evaluation [363] is necessary to improve the analysis and evaluation of anaphora resolution models, as well as to establish benchmarks for developing more accurate evaluation metrics. On the other hand, when designing new evaluation metrics, disagreements, and error severity should be considered by data-driven methods.

Model development. Regarding future advancements in anaphora resolution models, a significant area of focus should be on computational studies of anaphora resolution tasks that are firmly grounded in theory but have yet to be extensively explored. Examples of such tasks include but are not restricted to (1) bridging, deictic, and plural references, which are crucial aspects of referential language, yet their computational treatment has been limited, possibly due to a shortage of relevant annotated datasets; and (2) disagreement resolution, which involves learning from discrepancies in human interpretations of anaphoric expressions to better capture the pragmatic nuances of such references, and should be incorporated into future models [364]; and (3) cross-document anaphora resolution, which is critical for downstream applications such as knowledge graph construction and cross-document information extraction, yet has received insufficient attention in terms of data, methods, and evaluation metrics, particularly in relation to event resolution.

4. Named entity recognition

Name Entity Recognition (NER) is a critical component of Information Extraction, which involves identifying entity mentions in text, defining their boundaries, and assigning them entity types. The most commonly recognized entity types by NER systems are Location, Person, and Organization, and tokens referring to these entities are classified as entity mentions. In the following example:

(14) Steve Jobs is the founder of Apple.

an NER system would recognize the entities that “Steve Jobs” is Person; “Apple” is Organization. NER systems use pre-defined entity types, which may vary across different implementations. For example, Stanford’s widely used NER software [365] provides three versions that recognize three classes (Location, Person, Organization), four classes (Location, Person, Organization, Misc), and seven classes (Location, Person, Organization, Money, Percent, Date, Time), respectively. NER is a critical component in the field of NLP [366–368] and is often combined with other tasks, such as Relation Extraction (RE), to serve as a foundation for various NLP applications. Besides, NER is also used in various data mining tasks to recognize keywords, topics, and attributes [358,369,370].

NER can be traced back to the third Message Understanding Conference (MUC-3) [371]. The task for MUC-3 was designed to extract relevant information from the text and convert it into a structured format based on a predefined template, e.g., incident, the targets, perpetrators, date, location, and effects. Early NER systems that participated in MUC-3 primarily relied on rule-based approaches, which involved the manual creation of rules to identify named entities based on their linguistic and contextual features. However, with the dominance of deep learning in the NLP community, most NER tasks are now performed using neural networks. One of the first neural networks for NER was proposed by Collobert and Weston [372], which used a single convolutional neural network with manually constructed feature vectors. Later, this approach was replaced with high-dimensional continuous vectors, which were learned from large amounts of unlabeled data in an unsupervised manner [373]. With stronger models, now, the research in NER has been largely extended to nested NER [374], few-shot NER [375], joint entity and relation extraction (JERE)
[376,377].

Compared to standard NER whose entity relationship is absent, entities in nested NER have a hierarchical or nested structure, where one entity is embedded within another entity. For example, given

(15) The Ontario Supreme Court said ...

“Ontario” is a state entity that is embedded under the government entity of “Ontario Supreme Court” [378]. Given the very expensive annotation costs, few-shot NER is also a very important research trend. It learns NER with a limited amount of labeled data. JERE tasks are established based on the needs of downstream applications. In many cases, people not only need to know what an entity is but also need to know the relationship between entities. Thus, JERE needs to identify named entities in text as well as extract the relationships that exist between them. In the following example

(16) Greg Christie has been one of the greatest engineers at Apple.

For standard NER, “Greg Christie” should be identified as Person; “Apple” should be identified as Company. However, for JERE, besides the above entity recognition, an additional relationship label, “work_at” should also be predicted. Compared to identifying entities that are hierarchically structured within each other in nested NER tasks, the outcomes of JERE deliver another relationship dimension to connect entities. Both tasks are helpful in developing a comprehensive knowledge graph.

Due to the wide range of applications of NER, there have been several surveys conducted on this typical NLP task [379,380]. One recent study [381] focused specifically on NER in the biomedical field, also known as Bio-NER. In this domain, the presence of meaningless characters in biomedical data presents a significant challenge, particularly with regards to inconsistent word distribution. Similarly, [382] summarized and discussed the challenges specific to Chinese NER, rather than the more general English NER tasks. Meanwhile, [383] explored both NER and RE tasks, as they are closely linked and are typically composed of pipeline tasks. The aforementioned surveys focus on the technical perspective of NER, based on deep learning technology, while this section broadens the horizon of NER from theoretical foundations to applications.

4.1. Theoretical research

4.1.1. Prototype theory

Rosch [384] argued that our classification system, which includes the classification of named entities, is based on a central or prototype example. A prototype is a typical example of a category that represents the most common features or characteristics associated with the category. For example, the prototype of “bird” must associate the features, such as wings, feathers, and the ability to fly. Birds such as ostriches or penguins, which do not perfectly possess these characteristics, may be viewed as less typical examples. Rosch and Mervis [385] discovered that individuals can identify typical category examples faster and with greater precision than atypical examples. Thus, learning from prototypes can help to quickly grasp the important features of a named entity with a few examples.

4.1.2. Graded membership

Rosch et al. [386] argued that the classification of categories is frequently determined not by strict boundaries, but by various degrees of membership. We can use this theory for NER because the NER task also categorizes entities by predefined classes. The idea of Graded Membership implies how humans perceive and categorize the world around us. Some categories, e.g., “vegetable”, may be viewed as less distinct and vaguer. The theory suggests that the borders between categories may not be well-defined in some cases, leading to ambiguities when attempting to classify certain items, such as tomatoes or mushrooms. The ambiguity can be further compounded by cultural or regional differences in how categories are defined or classified.

4.1.3. Conceptual blending

According to [387], the act of blending different elements and their corresponding relationships is an unconscious process that is believed to be ubiquitous in everyday thought and language. This process involves the combination of various mental spaces or cognitive domains that are drawn from different scenarios and experiences. These scenarios may be derived from personal experiences, cultural practices, or societal norms, among others. Concept blending allows us to create a new concept by combining existing ones in novel ways. For example “SpaceX” may be mapped to mental spaces related to “aerospace” and “technology”; “Tesla” may be mapped to mental spaces related to “car” and “clean energy”. Conceptual blending provides an explanation for the recognition and comprehension of newly named entities by mapping them onto existing mental spaces or concepts.

4.1.4. Grammatical category

From the aspect of computational linguistics, the core issue of NER is how to define a named entity. Marrero et al. [388] group the criteria of a named entity as grammatical category, rigid designation, unique identification, and the domain of applications. However, many of the entity definitions in the NER domain are imperfect. From the view of grammatical category, a named entity is traditionally defined as a proper noun or a common name for a proper noun. Previous work has described NER as the recognition of proper nouns in general. However, as pointed out by Borrega et al. [389], the classic grammatical approach to proper noun analysis is insufficient to deal with the challenges posed by NER applications. For instance, in a toy question-answering task such as

(17) Do crocodiles live in the sea or on land?

“crocodiles”, “sea”, and “land” are not proper nouns, while they are commonly recognized as the essential entities for a proper understanding of the question. Consequently, a proper noun is no longer considered a criterion for identifying named entities in current NER research.

Table 10
The three common annotation schemes for NER.

Tokens:	West	African	Crocodile	are	semiaquatic	reptiles	that	live	in	Africa
IO	I	I	I	O	I	I	O	O	O	I
BIO	B	I	I	O	B	I	O	O	O	B
BIOES	B	I	E	O	B	E	O	O	O	S

Table 11
NER datasets and statistics.

Dataset	Source	# Sample	Reference
MUC-6	Newswire	318 articles	Grishman and Sundheim [392]
ACE-05	Social media	12,548 sentences	Walker et al. [395]
TACRED	Newswire	106,264 instances	Zhang et al. [396]
CoNLL-2003	Reuters[a]	1,499 articles	Sang and De Meulder [397]
I2B2	ECI Corpus[b]	1,600 patient records	Stubbs and Uzuner [398]
ADE	MEDLINE[c]	2,972 document	Gurulingappa et al. [399]
DDI	DrugBank[d]	1,025 document	Herrero-Zazo et al. [400]
WNUT-17	Social media	2,295 documents	Derczynski et al. [401]
OntoNote 5.0	Social media	–	Weischedel et al. [402]
CPR	MEDLINE	–	Krallinger et al. [403]
MultiNERD	Wikipedia	10 languages	Tedeschi and Navigli [404]
HIPE-2020	Newspapers	17,553 mentions	Ehrmann et al. [405]
NNE	Newswire	49,208 sentences	Ringland et al. [378]
GENIA	MEDLINE	18,546 sentences	Kim et al. [185]

[a] www.reuters.com/researchandstandards/

[b] http://www.ldc.upenn.edu/

[c] http://www.nlm.nih.gov/bsd/indexing/training/PUB_050.htm

[d] https://go.drugbank.com/

4.1.5. Rigid designation

The rigid designation is a concept in the philosophy of language which suggests that certain names or labels are inherently linked to the things they represent, e.g., “Barack Obama” rigidly designates the person who is the 44th President of the US, and it cannot be used to refer to any other person or entity. NER can be viewed as a form of rigid designation as it assigns labels to entities based on their intrinsic identity [390], rather than on their usage in the text. However, [391] noted that not all expressions that appear to designate rigidly can be analyzed as directly referring to an object in every possible world. This highlights the difficulty of defining entities with complex concepts in real-world applications. As a result, annotators likely make subjective judgments when labeling complex entities, which may be affected by entity descriptions and annotators’ understanding.

4.1.6. Unique identification

From the view of unique identification, the MUC conferences require that NER tasks annotate the “unique identification” of entities for all expressions [392]. However, determining what is unique depends on contextual elements, and can be a subjective process. While this “unique identification” is typically considered to be the reference being referred to, the definition itself poses a challenge in terms of defining what is truly unique.

4.1.7. Domain of applications

The definition of named entities was frequently grounded in the domain of applications. Entity definitions can be different between different NER tasks. For instance, in drug–drug interaction tasks [393], diseases may not be considered entities, whereas they are entities in adverse drug events [394]. Inconsistent entity definitions create challenges for machine learning. Because inconsistent entity definitions mean that for the same semantic unit, the machine has to summarize different entity representations to distinguish their labels under different tasks. This is also not conducive to training an all-around NER classifier on different application domains.

4.2. Annotation schemes

NER is typically approached as a sequence labeling task, where each token in a sentence is assigned a label. Three common annotation schemes are shown in Table 10. The IO scheme is a classification task that distinguishes between two classes, namely “Inner” and “Other”, to determine whether a token belongs to an entity or not. On the other hand, the BIO scheme employs three labels, namely “Beginning”, “Inner”, and “Other”, to identify tokens that represent the start of an entity, tokens that belong to an entity, and tokens that do not belong to any entity. The BIOES scheme expands on the BIO scheme by incorporating two additional labels, namely “Single” and “End”, to more precisely define the boundaries of entities.

By employing the IO scheme, the binary classification of tokens is simplified, as each token is labeled as either belonging to an entity or not. This straightforward labeling system makes it easier to identify entities in a text, but it fails to specify the position of the entities within the text. In contrast, the BIO scheme provides more precise annotations by identifying the beginning and continuation of an entity in the text. This labeling system allows for more accurate recognition of entities in a text and better classification of individual tokens. The BIOES scheme further extends the BIO scheme by providing more precise boundaries for entities, thereby allowing for better recognition of entity boundaries in a text. The “Single” label is used to denote an entity that consists of a single token, whereas the “End” label is used to indicate the final token of an entity. By incorporating these additional labels, the BIOES scheme provides a more nuanced approach to entity recognition and annotation.

4.3. Datasets

The surveyed popular NER datasets and their statistics can be viewed in Table 11. The first NER-focused dataset was published in the 6th MUC Conference [392]. This task consists of three sub-tasks, including entity names, temporal expressions, and number expressions. The defined entities include organizations, persons, and locations; The defined time expressions include dates and times; The defined quantities include monetary values and percentages.

More details can be seen in the office website.[21] The example of this dataset is shown as follows.

```
text: "Taga Co.",
type: "ORGANIZATION".
```

The MUC conference was replaced by Automatic Content Extraction (ACE) after 1997. ACE05 [395] is another popular NER dataset published at ACE Conference. ACE05 is a multi-lingual dataset, which contains English, Arabic, and Chinese data. The corpus consists of data of various types annotated for entities, relations, and events. Its data source includes broadcast conversation, broadcast news, newsgroups, telephone conversations, and weblogs. More details can be seen on the office website.[22] The example of this dataset is shown as follows.

```
entity id: "NN_ENG_20030630_085848.18-E1",
type: "GPE",
subtype: "State-or-Province",
class: "SPC",
start: "82",
end: "91",
name: "california".
```

After MUC, the Text Analysis Conference (TAC) published the Knowledge Base Population challenge. In this challenge, the Stanford NLP Group developed TAC Relation Extraction Dataset (TACRED) [396], which contains 106,264 instances with annotated entities, relations and some other NLP tasks. More details can be seen on the office website.[23] The example of this dataset is shown as follows.

```
id: "e7798fb926b9403cfcd2",
docid: "APW_ENG_20101103.0539",
relation: "per:title",
token: "[‘At’, ‘the’, ‘same’, ‘time’, ‘,’, ‘Chief’, ...]",
subj_start: "8",
subj_end: "9",
obj_start: "12",
obj_end: "12",
subj_type: "PERSON",
obj_type: "TITLE",
stanford_pos: "[‘IN’, ‘DT’, ‘JJ’, ‘NN’, ‘,’, ‘NNP’, ‘NNP’, ...]",
stanford_ner: "[‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ...]"
stanford_head: "[4, 4, 4, 12, 12, 10, 10, 10, 10, 12, ...]",
stanford_deprel: "[‘case’, ‘det’, ‘amod’, ‘nmod’, ‘punct’,
....]".
```

CoNLL-2003 [397] is another widely used NER dataset. This task concerned language-independent named entity recognition, which concentrates on four kinds of named entities: locations, persons, organizations, and names of miscellaneous entities that do not belong to the previous three kinds. The related data files are available in English and German. More details can be seen on the office website.[24] The example of this dataset is shown as follows.

```
text: "[‘U.N.’, ‘official’, ‘Ekeus’, ‘heads’, ...], ",
pos: "[‘NNP’, ‘NN’, ‘NNP’, ‘VBZ’, ...], ",
syntactic chunk: "[‘I-NP’, ‘I-NP’, ‘I-NP’, ‘I-VP’, ...], ",
named entity tag: "[‘I-ORG’, ‘O’, ‘I-PER’, ‘O’, ...]".
```

Table 12
Useful knowledge bases for NER.

Name	Knowledge	# Entities	Structure
Wikipedia	World	13,489,694	Unstructured
Wikidata	World	100,905,254	Graph
DrugBank	Medical	over 500,000	Structured
UMLS	Medical	16,857,345	Structured
BioModels	Medical	unclear	Structured
SNOMED CT	Medical	over 350,000	Structured
ICD-10	Medical	unclear	Structured
MIMIC-III	Medical	unclear	Structured
MeSH	Medical	over 28,000	Structured
GeoNames	Geographical	over 25,000,000	Structured
EDGAR	Financial	unclear	Structured
EduKG	Educational	5,452	Structured

Besides the above famous datasets, MultiNERD [404], HIPE-2020 [405], and NNE [406] are also popular NER datasets in general domain. NER tasks have garnered considerable attention in numerous specialized domains. Informatics for Integrating Biology and the Bedside (I2B2) [398] is a national biomedical computing project sponsored by the National Institutes of Health (NIH) from 2004 to 2014. I2B2 actively advocates mining medical value from clinical data and has organized a series of evaluation tasks and workshops for unstructured medical record data, and these evaluation tasks and open datasets have gained wide influence in the medical NLP community. I2B2 is maintained in the Department of Biomedical Information at Harvard Medical School and continues to conduct assessment tasks and workshops, and the project has been renamed National NLP Clinical Challenges (N2C2). More details can be seen on the office website.[25] Besides, there also exist many other biomedical datasets for specific medical NER tasks, including Adverse Drug Events (ADE) [399,407], Drug–Drug Interaction [400], and Chemical Protein Reaction (CPR) [403], and GENIA [408].

4.4. Knowledge bases

Table 12 illustrates useful knowledge bases for NER. The biggest ones are Wikidata,[26] and Wikipedia[27] which are multi-lingual free online encyclopedias maintained by worldwide volunteers.

There are also knowledge bases in a specific field. SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) [409] is a systematically organized collection of medical terms that provides a standardized representation of clinical information, which is often used in NER tasks involving clinical data. MeSH (Medical Subject Headings) [410] is another controlled vocabulary, developed by the U.S. National Library of Medicine. It is used for indexing and organizing biomedical literature. Other medical knowledge bases include UMLS (Unified Medical Language System) [411,412], ICD-10 [413], MIMIC-III [414], DrugBank [415], and bioinformatics knowledge base BioModels [416]. GeoNames [417] is a comprehensive geographic knowledge repository that encompasses over 25 million geographical names and comprises over 11 million distinctive features, including cities, countries, and landmarks. EDGAR (Electronic Data Gathering, Analysis, and Retrieval) [418] is a database maintained by the U.S. Securities and Exchange Commission (SEC), containing financial filings and reports from publicly traded companies. EduKG [419] is an educational knowledge base.

4.5. Evaluation metrics

In the process of named entity recognition task evaluation, the main evaluation metrics are also Precision, Recall, and F-value.

[21] https://cs.nyu.edu/~grishman/NEtask20.book_2.html
[22] https://catalog.ldc.upenn.edu/LDC2006T06
[23] https://nlp.stanford.edu/projects/tacred/#intro
[24] https://www.clips.uantwerpen.be/conll2003/ner/
[25] https://www.i2b2.org/
[26] https://www.wikidata.org/
[27] https://en.wikipedia.org/

4.6. Annotation tools

One AI[28] is an online platform that offers NLP-as-a-service. The utilization of APIs enables developers to effectively analyze, manipulate, and transform natural language inputs within their programming code without requiring any specialized knowledge of NLP. One AI facilitates the interpretation of both the meaning and information conveyed in textual data, and can produce structured data in context via language processing.

GATE Teamware[29] [420] is an integrated annotation tool for comprehensive language processing tasks, especially for Information Extraction systems. The University of Sheffield developed GATE Teamware that enables collaborative semantic annotation projects through a shared annotation environment. The software comprises several beneficial attributes such as the ability to load document collections, create project templates that can be used multiple times, initiate projects based on templates, assign project roles to individual users, monitor progress and obtain various project statistics in real-time, report project status, annotator activity, and statistics, and apply automatic annotations or post-annotation processing via GATE-based processing routines.

MAE[30] [421] (Multi-document Annotation Environment) is a general-purpose and lightweight natural language annotation tool. The tool enables users to specify and create their customized annotation tasks, annotate any text spans of their choice, utilize non-consuming tags, effortlessly establish links between annotations, and produce annotations in stand-off XML format. It also provides a simple adjudication process with a visualization feature that displays the extent tags, link tags, and non-consuming tags of any XML standoff annotated documents.

UIMA[31] [422] (Unstructured Information Management Applications) is a framework that falls under the purview of the Apache Software Foundation. It serves as a comprehensive platform for managing language processing projects and is licensed under Apache's open-source license. With its versatile capabilities, UIMA can effectively handle a diverse array of language processing tasks and extract various types of information. The UIMA's Regular Expression Annotator is capable of identifying entities such as email addresses, phone numbers, URLs, zip codes, or any other entities based on the utilization of regular expressions and concepts. The tool can generate an annotation for each detected entity or update an existing annotation with relevant feature values.

Brat[32] (Browser-based Rapid Annotation Tool) is a free data labeling tool that offers a seamless browser-based interface for annotating text. It streamlines numerous annotation tasks related to natural language processing. With a thriving support community, Brat is a well-known and widely used tool in NER. It also offers the option of integrating with external resources, such as Wikipedia. Moreover, Brat enables organizations to establish servers that allow multiple users to collaborate on annotation tasks. However, implementing this feature does necessitate some technical proficiency and server management skills.

4.7. Methods

4.7.1. Nested NER

A. Multi-label Method

Due to the fact that nested named entities can have multiple labels for a single token, traditional sequence labeling methods are not directly applicable to the recognition of nested named entities. To address this issue, researchers have attempted to convert the multi-label problem into a single-label problem or adjust the decoder to assign multiple labels to the same entity.

Katiyar and Cardie [423] proposed a method to address nested named entity recognition by modifying the label representation in the training set. Instead of using one-hot encoding, they used a uniform distribution over the specified classes as the label. During inference, a hard threshold is set and any class with probability above this threshold is predicted for the token. However, this approach has two limitations: it is difficult to determine the objective for model learning; the method is sensitive to the manually chosen threshold value.

Straková et al. [424] changed nested NER from multi-label to single-label tasks by modifying the annotation schema. They combined any two categories that may co-occur to produce a new label (e.g., combine B-Location with B-Organization to construct a new label B_Loc_Org). One benefit of this approach is that the final classification task is still a single category because all possible classification targets had been covered in the schema. Nonetheless, this method brought about a proliferation of label categories in an exponential manner, leading to sparsely annotated labels that proved difficult to learn, particularly in the context of entities nested across multiple layers.

In order to address the issue of label sparsity, [408] proposed a hierarchical approach. If the classification of nested entities cannot be resolved in a single pass, the classification is continued iteratively until either the maximum number of iterations is reached or no new entities can be generated. Nevertheless, this approach is susceptible to error propagation, whereby an erroneous classification in a preceding iteration could impact subsequent iterations.

B. Generation-based Method

Li et al. [425] proposed a unified framework to accomplish flat and nested NER tasks by formulating NER as a machine reading comprehension (MRC) task [426]. In this approach, the extraction of each entity type corresponds to specific questions. For instance, when the model is given the question "which location is mentioned in the sentence?" along with the original sentences, it generates an answer such as "Washington". This approach is similar to Prompt Tuning [427], which avoids the labor-intensive process of constructing manual questions. However, in this method, the generated tokens must be mapped to pre-defined named entity types.

Yan et al. [428] proposed a novel pointer generation network. Given an input sentence, the model generates the entity indexes in this sentence that belong to entities. In such a way, flat, nested, and discontinuous entities can be recognized in a unified framework. Skylaki et al. [429]; Fei et al. [430]; Yang and Tu [431]; Su et al. [374] are also following the idea of generating indexes of a sentence to recognize nested entities.

C. Hypergraph-based Method

A hypergraph is a generalized variant of a normal graph, which is characterized by an edge that can connect an arbitrary number of vertices [432]. It is widely used in the NLP community for the tasks of syntactic parsing, semantic parsing, and machine translation because it can accurately describe the relationship between objects with multiple associations. A set of objects with only binary relations can be described by a normal graph. However, when the objects are often related to each other in a more complex one-to-many or many-to-many, e.g., nested named entities, hypergraphs become a more appropriate data structure. A typical example of nested NER with a hypergraph solution is shown in Fig. 4. Finkel and Manning [433] firstly introduced hypergraphs into nested NER tasks, named Mention Hypergraph. In their model, Mention Hypergraph utilized nodes and directed hyper-edges to jointly represent named entities and their combinations. To compute the training loss, the proportion of accurate structures was calculated and divided by a normalized term. This term was obtained using a dynamic programming algorithm that aggregated feasible nested sub-graphs for NER. However, the normalized terms obtained from this algorithm included fractions of pseudo-structures, which led to errors. To deal with the problem of pseudo-structures, [434] proposed a gap-based marker model to identify nested entity structures by combining mention separators with features. In this method, the authors manually designed 8 types of mention separators for various scenarios.

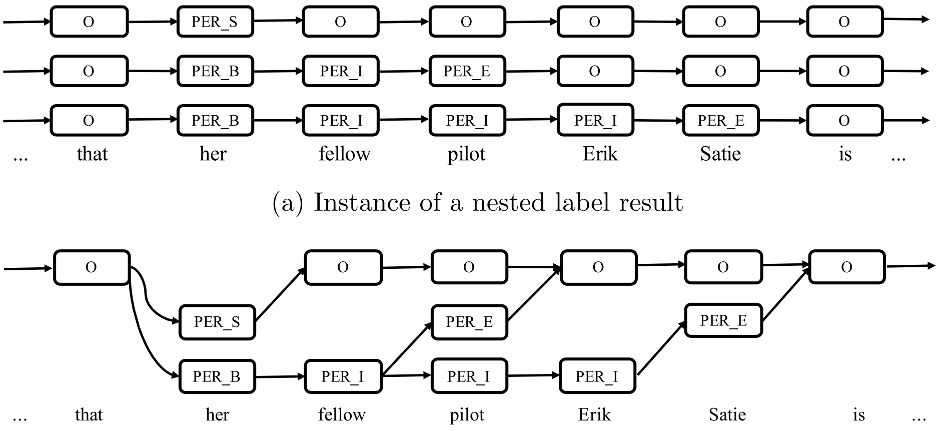
[28] https://docs.oneai.com/docs

[29] https://gate.ac.uk/teamware/

[30] https://keighrim.github.io/mae-annotation/

[31] https://uima.apache.org/sandbox.html

[32] https://brat.nlplab.org/



(a) Instance of a nested label result

(b) Corresponding hypergraph structure

Fig. 4. A typical example for nested NER with hypergraph solution.

Based on the mention separators' states for any two consecutive tokens, they defined accurate and novel graph structures. However, since this approach only utilized local information to construct the graph structures, it may not be unambiguous for long-nested named entities. For instance, when presented with the nested entity "a West African Crocodile", which includes two separate entities, "West African" and "a West African Crocodile", their approach may also recognize "a West African" as a named entity. This ambiguous problem was solved by Wang and Lu [435], which proposes a segmental hypergraphs method. The method used an unambiguous ambiguity-free compact hypergraph representation to encode all possible combinations of nested named entities. Upon Mention Hypergraph [433], segmental hypergraphs employed an inside–outside message-passing algorithm that can summarize the features of child nodes to the parent node and achieve efficient interference. Besides the above work, [436] introduced the concept of regional hypernodes and a combination method of graph convolutional network (GCN) and BiLSTM to generate hypernodes for each region. Yan and Song [437] employed start token candidates and generated corresponding queries with related contexts, then used a query-based sequence labeling module to form a local hypergraph for each candidate.

4.7.2. Few-shot NER

A. Metric Learning

Metric Learning is a common technology in various few-shot tasks. Prototypical Networks [438] is a milestone in few-shot metric learning. Prototypical Networks compute the centroid of each category based on the support set. They determine the distance between the samples in the query set and the prototype center, followed by updating the model by optimizing this distance. Upon completion of the training phase, the embedding of each sample will be situated in closer proximity to the centroid of the corresponding category. Such an idea was largely inspired by Prototype Theory (see Section 4.1.1). Fritzler et al. [439] adopted the prototypical network into few-shot NER tasks. They argued that words in a sentence are interdependent and, therefore, the labeling of adjacent words should be taken into account. To address this issue, they substituted the conventional token input of Prototypical Networks with complete sentences. However, this method ignores the problem of the Outside (O) class in NER tasks, which actually represent different semantic meanings. This problem would significantly affect the model's performance under few-shot settings.

To avoid the above issues, [440] followed the nearest neighbor inference [441] to assign labels to tokens. In contrast to Prototypical Networks, which learn a prototype for each entity class, this study characterized each token by its labeled instances in the support set alongside its context. The approach determined the nearest labeled token in the support set, followed by assigning labels to the tokens in the query set that require prediction. Das et al. [442] proposed CONTaiNER, which optimized the inter-token distribution distance. CONTaiNER employed generalized objectives to different token categories based on their Gaussian-distributed feature vectors. Such a method has the potential to mitigate overfitting problems that arise from the training domains.

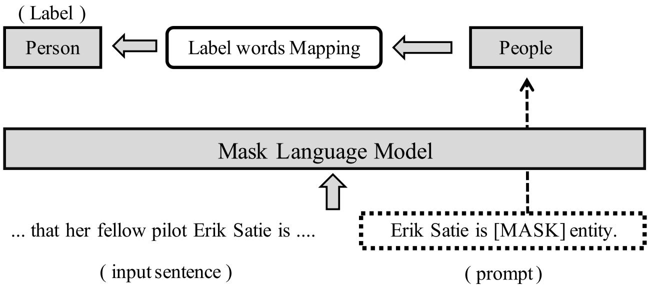
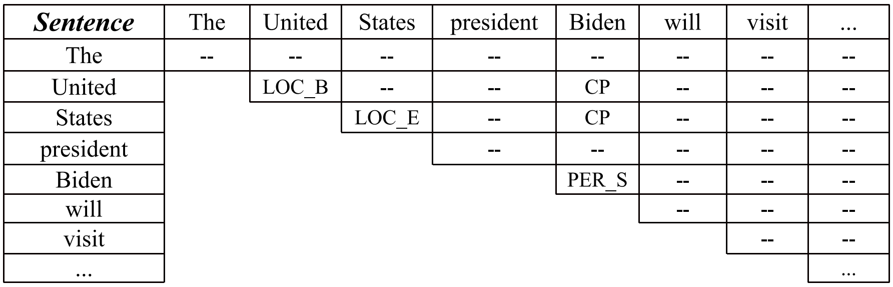


Fig. 5. A typical prompt tuning example for NER tasks.

B. Prompt Tuning

Recently, prompt tuning has shown great potential on few-shot tasks by reformulating other tasks as mask language tasks [443–445]. Prompt tuning-based methods need construct prompts to obtain masked word predictions and then map predicted works into pre-defined labels, as shown in Fig. 5. Cui et al. [446] proposed a template-based method for NER, which first applied the prompt tuning to NER tasks. However, their method had to enumerate all possible spans of sentences combined with all entity types to predict labels, which suffered serious redundancy when entity types or sentence lengths increased. Manually defined prompts were labor-intensive and made the algorithm sensitive to these prompts. To avoid the manual prompt constructions, [447] tried to explore a prompt-free method for few-shot NER. The present study introduced an entity-oriented language model that decodes input tokens into their corresponding label words if they belong to entities. In cases where the tokens are not entities, the entity-oriented language model decodes the original tokens. Nevertheless, this approach encounters difficulties in labeling word engineering. While this study proposed an automated label selection technique, the associated experiments revealed some degree of instability. COPNER [448] introduced class-specific words to construct prompt tuning. By comparing each token with manually selected class-specific words, this method needed neither manual prompts nor label words engineering. The selected class-specific words (a representative word corresponding to a class)



Sentence	The	United	States	president	Biden	will	visit	...
The	--	--	--	--	--	--	--	--
United		LOC_B	--	--	CP	--	--	--
States			LOC_E	--	CP	--	--	--
president				--	--	--	--	--
Biden					PER_S	--	--	--
will						--	--	--
visit							--	--
...								...

Fig. 6. The illustration of the table-filling strategy.

were directly concatenated with original sentences as prompts. However, the manual selection of class-specific words is subjective, and a single word may not entirely capture the semantics of an entity category.

4.7.3. Joint NER and relation extraction

A. Parameter Sharing-based Multi-tasks Learning

Considering that NER is usually combined with relation extraction tasks applied in various downstream tasks, jointly recognizing named entities and classifying relations is a hot topic in related fields. Multi-task learning is the most common solution in joint NER and relation extraction. Miwa and Bansal [449] firstly employed a shared Bi-LSTM encoder to obtain token representations, and then fed encoded representations into NER and relation extraction classifiers, respectively. Sun et al. [450] utilized a GCN as a shared encoder to enable joint inference of both entity and relation types. The core idea of the above study is that multi-task models can enhance the interactions between the learning of NER and relation extraction, and further alleviate the error propagation by sharing common parameters [368]. However, this work cannot ensure that the sharing of information is useful and proper. NER and relation extraction might need different features to result in precise predictions.

To deal with such a problem, [451] proposed an information filtering mechanism to provide valid features for NER and relation extraction. Their method used an entity and relation gate to divide cell neurons into different parts and established a two-way interaction between NER and relation extraction. In the final employed network, each neuron contained a shared partition and two task-specific partitions.

B. Table Filling

While multi-task learning can improve the interdependence between NER and relation extraction, the relation extraction process still requires the pairing of all entities from the NER tasks to classify relations, making it impossible to completely eliminate error propagation. To solve the problem, [452] proposed a table-filling strategy to achieve joint NER and relation extraction by labeling input tokens in a table. The method utilized token lists of sentences to form rows and columns. Then, they extracted entities using the diagonal elements and classified relations with a lower/upper triangular matrix of the table. This basic table-filling strategy can be seen in Fig. 6. Nonetheless, this approach involved the explicit integration of entity-relation label interdependence, which necessitated the use of intricate features and search heuristics.

Gupta et al. [453] incorporated neural networks with a table-filling strategy via a unified multi-task recurrent neural network. This method detected both entity pairs and the related relations with an entity-relation table, which alleviated the need for search heuristics and explicit entity-relation label dependencies. Zhang et al. [454] further integrated global optimization and syntax information into the table-filling strategy to combine NER and relation extraction tasks. Ren et al. [455] argued that the above table-filling-based studies only focus on utilizing local features without the global associations between relations and pairs. Ren et al. [455] first produced a table feature for every relation, followed by extracting two types of global associations from the generated table features. Finally, the table feature for each relation was integrated with the global associations. Such a process is performed iteratively to enhance the final features for joint learning of NER and relation extraction tasks.

C. Tagging Scheme

The table-filling approach can mitigate issues related to error propagation. However, these techniques require the pairing of all sentence elements to assign labels, resulting in significant redundancy. To address the redundancy and avoid error propagation, [456] proposed a novel tagging scheme that converted joint NER and relation extraction into a united task. The idea was similar to the solution for nested entities [424], which combined NER labels with relation extraction labels by modifying the annotation schema. For example, given the sentence “The United States president Biden will visit ...”, by allocating the customized labels “Country-President_B_1”, “Country-President_E_1” for tokens “United”, “States”, and “Country-President_E_2” for token “Biden”, the proposed method can directly obtain the triplet (United State, Country-President, Biden).

Yu et al. [457]; Wei et al. [458] proposed two similar methods. In contrast to conventional joint approaches for NER and relation extraction, which involve recognizing entities followed by relation classification, the two methods first identified all head entities. Next, for each identified head entity, they simultaneously predicted corresponding tail-entities and relations, achieving cascade frameworks combined with a customized tagging scheme. The typical joint NER and relation extraction tasks learn to model the conditional probability:

$$P(h,r,t) = P(s)P(t \mid h)P(r \mid h,t), \tag{7}$$

where h represent head entity; r represent relation; h represent tail entity. The above methods combined the last two parts in Eq. (7), yielding

$$P(h,r,t) = P(s)P(t,r \mid s). \tag{8}$$

4.8. Downstream applications

4.8.1. Knowledge graph construction

Knowledge graphs are structured semantic knowledge bases for rapidly describing concepts and their interrelationships in the physical world, aggregating large amounts of knowledge by reducing the data granularity from the document level to the instance level [459]. Thus, knowledge graphs enable rapid response and reasoning about knowledge. At present, the application of knowledge graphs has become prevalent in industrial domains, such as Google search. Generally, the construction of Knowledge Graphs consists of three main parts: information extraction, information fusion, and information processing. The task of information extraction involves the identification of nodes through NER and the establishment of edges via relation extraction. The task of information fusion is utilized for normalizing nodes and edges. The normalized nodes and edges need to go through a quality assessment with the task of information processing to be added to knowledge graphs.

He et al. [368] proposed a multi-task learning-based method for the construction of genealogical knowledge graphs. At first, [460] collected unstructured online obituary data. Then, they extracted named entities as nodes and classified family relationships for these recognized people as edges to construct genealogical knowledge graphs. Similarly, [461] utilized NER and relation extraction for obtaining the nodes and edge in biomedical knowledge graphs. They proposed a customized tagging schema to convert the construction of biomedical knowledge graphs into a sequence labeling task with multiple inputs and multiple outputs. Li et al. [462] proposed a systematic approach for constructing a medical knowledge graph, which involves extracting entities such as diseases and symptoms, as well as related relationships, from electronic medical records. Silvestri et al. [463], Peng et al. [464], and Shafqat et al. [465] aimed to collect and utilize medical knowledge for NER. Further, constructing knowledge graphs requires the task of Entity Linking [466] to normalize entities with different names. Entity Linking and NER are typically performed as pipeline tasks to yield more nodes and edges for the constructed graphs. Additionally, Entity Linking can be seen as a downstream task for NER, as it further refines the identified entities by linking them to a specific reference entity in a knowledge graph.

4.8.2. Recommendation systems

Recommendation systems can be classified into two primary categories based on their solutions, namely content-based recommenders and collaborative filtering-based recommenders [467]. For both of these groups, gathering data on users and products is a crucial step in the entire process. In this regard, NER modules play a pivotal role. For example, [468] introduced the 5W1H model, which utilizes NER to extract contextual information, specifically Who, Why, Where, What, When, and How, to generate contextual recommendations.

Zhou et al. [469] argued that recommendation systems currently in use suffer from a deficiency of contextual information in conversational data, as well as a semantic gap between natural language expressions and the preferences of individual users for specific items. To overcome these challenges, word- and entity-oriented knowledge graphs were incorporated to enhance the data representations. Mutual Information Maximization was adopted to align the word-level and entity-level semantic spaces. The aligned semantic representations were used to develop a knowledge graph-enhanced recommender component to make accurate recommendations, and a knowledge graph-enhanced dialog component that can generate informative keywords or entities in the response text. A NER module is a crucial component in creating such a knowledge graph-enhanced system [470].

Iovine et al. [471] proposed a domain-independent, configurable recommendation system framework, named ConveRSE (Conversational Recommender System framEwork). ConveRSE utilized various interaction mechanisms, including natural language, buttons, and a combination of the two. The framework comprised a dialog manager, an intent recognizer, a sentiment analyzer, an entity recognizer, and a set of recommendation services. The entity recognizer component specifically focused on identifying relevant entities that were mentioned in the user's input, and linking them to an appropriate concept in the knowledge base. The ConveRSE framework's success is heavily reliant on the performance of the NER component, as it plays a crucial role in enhancing the system's overall performance.

Wang et al. [472] proposed RippleNet, an end-to-end framework that incorporates the knowledge graph into a recommender system. RippleNet overcame the limitations of previous embedding-based and path-based approaches to knowledge graph-aware recommendation by incorporating the knowledge graph as a form of supplementary information. RippleNet included both inward aggregation and outward propagation models. The inward aggregation version aggregated and incorporated neighborhood information when computing the representation of a given entity. By extending the neighborhood to multiple hops away, it was possible to model high-order proximity, thereby capturing users' long-distance interests. On the other hand, the outward propagation model propagated users' potential preferences and explored their hierarchical interests in knowledge graph entities.

Upadhyay et al. [473] proposed an explainable job recommendation system by matching users with the most pertinent jobs, based on their profiles. The system also provided a human-readable explanation for each recommendation. The NER module was customized to extract pertinent details from both job postings and user profiles. These details were utilized to create comprehensible explanations for each recommendation. By identifying and categorizing entities, the NER module enhanced the accuracy and understandability of the textual explanations, providing a clear representation of the reasoning behind the recommendation system.

4.8.3. Dialogue systems

Commonly, dialogue systems can be categorized into three main types, namely task-oriented, question-answering, and open-domain [474]. NER plays a role in enhancing the natural language understanding of the three types of dialogue systems, organizing original user messages into semantic slots, and classifying data domain and user intention [475]. Abro et al. [476] proposed an argumentative dialogue system with NER and other natural language understanding tasks. The approach can enhance comprehension of user intent by comprehending injected entities and relationships. For the question-answering [477] and open-domain dialogue systems, NER also plays a crucial role in the part of intent recognition and knowledge retrieval. For example, [478] developed a sequence of sub-goals with external knowledge to improve generation performance. External knowledge refers to a range of named entities and relationships that are associated with a conceptual entity. Leveraging external knowledge allows the dialogue system to deliver a more cohesive small talk from the open domain.

4.9. Summary

NER is a very important semantic processing technique for information retrieval. It is the manifest of cognitive semantics, because named entities are not simply categorized by their semantics. The classified named entities also reflect their inherent attributes in people's cognition. According to Prototype Theory (see Section 4.1.1), the inherent attributes of named entities can be represented by prototypes. It is gratifying to observe that a theory has had a significant influence on research related to few-shot NER. On the other hand, the ambiguity of named entity classification argued by Graded Membership (see Section 4.1.2) and Grammatical Category (see Section 4.1.4) was rarely analyzed from computational linguistic aspects. We also do not see explainable NER studies that explain why an entity is classified into a particular category from the perspective of conceptual blending (see Section 4.1.3). The NER research on these aspects is helpful for achieving human-like intelligence in categorizing named entities.

The availability of numerous named entity recognition (NER) datasets, both in general and medical domains, has significantly enhanced computational research in this area. This may be attributed to the great application value of NER, as well as a wide range of data annotation tools. Encyclopedias knowledge and domain-specific knowledge also provide external information to help NER models better understand the context and commonsense. Now, NER has developed many practical task setups to the need of technical applications, e.g., nested NER, few-shot NER, joint NER and relation extraction, and downstream tasks, e.g., knowledge graph construction, recommendation systems, and dialogue systems.

Table 13
A summary of representative NER techniques. The study with * means it cannot be compared with other studies since it did not report 5-shot results.

Task	Reference	Tech	Feature and KB.	Framework	Dataset	Score	Metric
Nested NER	Katiyar and Cardie [423]	DL	Emb.	Bi-LSTM	ACE-05	70.2%	F1
	Straková et al. [424]	DL	Emb.	LSTM-CRF	ACE-05	84.3%	F1
	Shibuya and Hovy [408]	DL	Emb.	LSTM-CRF	ACE-05	84.3%	F1
	Li et al. [425]	DL	BERT, Wikipedia	Unified framework	ACE-05	86.9%	F1
	Yan et al. [428]	DL	BERT	Pointer networks	ACE-05	84.7%	F1
	Finkel and Manning [433]	Graph	Emb., Constituency parsing	Hypergraph	GENIA	72.0%	F1
	Muis and Lu [434]	Graph	Emb., Multigraph representation	Hypergraph	GENIA	70.8%	F1
	Wang and Lu [435]	Graph	Emb., Segmental hypergraphs	Hypergraph	GENIA	75.1%	F1
	Yang and Tu [431]	DL	BERT	Pointer networks	ACE-05	85.0%	F1
	Su et al. [374]	DL	BERT	Pointer networks	CONLL04	88.6%	F1
Few-shot NER (5 shot)	Fritzler et al. [439]*	DL	Prototypical network	RNN+ CRF	Ontonotes	–	F1
	Yang and Katiyar [440]	DL	BERT	Nearest neighbor	I2B2	22.1%	F1
	Das et al. [442]	DL	BERT	Contrastive learning	I2B2	31.8%	F1
	Cui et al. [446]	DL	BERT	Prompt tuning	I2B2	36.7%	F1
	Huang et al. [375]	DL	BERT	Prompt tuning	I2B2	43.7%	F1
Joint NER and RE	Miwa and Bansal [449]	DL	Emb.	Bi-LSTM	ACE-05	55.6%	F1
	Sun et al. [450]	Graph	Emb.,	Bipartite graph	ACE-05	59.1%	F1
	Yan et al. [451]	DL	BERT	Partition filter	ACE-05	66.8%	F1
	Gupta et al. [453]	ML	Emb.,	Table filling	CoNLL04	72.1%	F1
	Zhang et al. [454]	DL	Emb.,	Table filling	ACE-05	57.5%	F1
	Zheng et al. [456]	DL	Emb.,	Tagging scheme	NYT	49.5%	F1
	Yu et al. [457]	DL	Emb.,	Tagging scheme	NYT	59.0%	F1
Task-driven NER	Shafqat et al. [465]*	DL	Emb., ICD-10	Bi-LSTM	No public	–	F1
	Hirsch et al. [413]*	DL	Emb., UMLS	Bi-LSTM	No public	–	F1
	Peng et al. [464]*	DL	BERT, MIMIC-III	Fine tuning	No public	–	F1

4.9.1. *Technical trends*

Due to the extensive research conducted on typical NER methods over the years, researchers are shifting their focus towards NER techniques that are more applicable to practical scenarios, for example, nested NER, few-shot NER, and joint NER and relation extraction. Recent technological trends for the aforementioned NER tasks are summarized in Table 13. Overall, nested NER can be addressed by multi-label, generation-based, and hypergraph-based methods. Among them, multi-label methods are straightforward and easy to implement. However, there are several limitations in the surveyed multi-label methods. For example, thresholds for multi-label selection are hard to decide empirically [423]; multiple labels are suffering sparsity [424] or error propagation [408], which can lower model performance. Generation-based methods are flexible. By reformulating NER tasks as question-answering, they can generate any results which satisfied the pre-defined requirements [408,425]. These methods are used for handling Flat NER [429], nested NER [428], and discontinuous NER [430]. However, a generation-based method is hard to control what is generated, even if some studies [374,429–431] have attempted to restrict the outputs of generation-based methods to a specific set of indexes (pointer network). The core point of the hypergraph-based method is about how to establish a hypergraph data structure to better represent interaction among all tokens in a sentence. These methods are good at modeling the interactions among all tokens in a sentence. It is important to note that the majority of hypergraph-based methods exhibit a task-specific nature, indicating a limited scope of applicability. These methods may not be universally applicable, and their effectiveness may be constrained by the specific task they are designed for.

Few-shot NER is usually achieved by metric learning and prompt tuning. Metric learning has demonstrated its effectiveness in various few-shot tasks [439,479]. For few-shot NER tasks, some works predict the final labels by comparing token-to-token distance [440,442] or token-to-prototype distance [448]. These methods have to decide different distance calculation functions according to different task [480] and suffer instability introduced by insufficient data. By exploiting the full potential of language models, prompt tuning is proposed and demonstrated as a promising technology for few-shot tasks [427,443,481]. Prompt tuning reformulate NER as a mask language model task to reduce the gap between NER and employed pre-training LMs. The drawback is that prompt tuning needs extra template construction and label word mappings and some studies have tried to deal with such problems [448]. For Joint NER and RE tasks, we summarize related studies into three groups, including parameter sharing-based multi-task learning, table-filling strategy, and customized tagging scheme. Parameter sharing is the basic idea in multi-task learning, which can be used to enhance the interaction between NER and RE [482,483]. This method can provide some relief from error propagation, but it cannot completely eliminate the issue. Also, this method has to pair every two entities for relation extraction, which introduces unnecessary redundancy. Table filling-based joint NER and relation extraction can completely eliminate error propagation by converting NER and relation extraction into a whole sequence-tagging task [453,455,484]. However, these methods have to label every two token pairs in an input sentence in an enumerable fashion. If relation extraction is defined as an unidirectional task, the half of calculations are wasted. Following the idea of the table filling strategy, tagging scheme-based methods also model the NER and relation extraction as an integrated task. The fundamental concept of the tagging scheme is to merge the labels assigned for NER with those assigned for relation extraction into a unified label [424,456]. Such a method has the potential to circumvent issues related to both error propagation and redundancy; however, it may also lead to a sparsity of labels.

4.9.2. *Application trends*

We have discussed three main downstream applications of NER, including knowledge graph construction, dialogue systems, and recommendation systems. Table 14 illustrate related studies. Usually, NER is the basic module for providing recognized entities for further utilization. In this case, a NER model works as a parser to mine knowledge from unstructured text. The recognized entities and relations can be used as nodes and edges for knowledge graph construction. The entities can also serve as intent recognition methods in recommendation systems, and slot-filling methods in dialogue systems. For example, [486] proposed a pre-trained task-oriented dialogue BERT, which significantly boosts the performance of a dialogue system by improving the intent detection sub-task. Wang et al. [487] proposed a method for recognizing related spans and value normalization with slot attention to improve the dialogue system. Besides, we also observe that using the identified named entities as features can also improve the performance

Table 14
A summary of the representative applications of NER in downstream tasks. ✓ denotes the role of NER in a downstream task.

Reference	Downstream task	Feature	Parser	Explain.
Yao et al. [459]	Knowledge graph construction		✓	
He et al. [368]	Knowledge graph construction		✓	
Jiang et al. [461]	Knowledge graph construction		✓	
Li et al. [462]	Knowledge graph construction		✓	
Kim et al. [468]	Recommendation systems		✓	✓
Adomavicius and Tuzhilin [485]	Recommendation systems	✓		
Zhou et al. [469]	Recommendation systems	✓		
Iovine et al. [471]	Recommendation systems	✓		
Wang et al. [472]	Recommendation systems		✓	
Li et al. [475]	Dialogue systems		✓	
Abro et al. [476]	Dialogue systems		✓	
Dimitrakis et al. [477]	Dialogue systems		✓	
Zhang et al. [478]	Dialogue system		✓	✓

of recommendation systems, because NER can help identify important entities that could be useful for making recommendations. The most common problem is error propagation between NER and other components in a downstream system. Kim et al. [488] employed a two-step neural dialog state tracker to alleviate the impact of the original error. With the development of PLMs and LLMs, many downstream tasks are organized as end-to-end processing tasks to achieve higher accuracy and mitigate error propagation issues. However, we can still observe that NER can improve the explainability in recommendation and dialogue systems [468,478], which is also an important aspect of AI research. There is still a considerable untapped potential for integrating NER with other downstream tasks, e.g., explaining how concepts blend each other between different entities; what the inherent attribute of a group of entities the selected prototypes represent; how robust an identified named entity is.

4.9.3. Future works

Open-domain NER. Compared with typical single-domain NER, open-domain NER has more categories. Besides, the entity classes are hardly defined in advance. For such reason, open-domain NER is more capable of handling rapidly expanding data, and mining more potential knowledge which is hidden in massive unstructured text data [489,490]. Open-domain NER is significant because it discovers and connects world knowledge via automatic text mining. Many manually developed lexical resources, e.g., WordNet can only cover limited concepts. When the concepts come to multi-word expressions, manually mining, structuring and updating those concepts can result in the exponential growth of human efforts. Open-domain NER is helpful for mitigating human efforts and delivering a knowledge base that connects entities from different domains.

Multi-lingual NER. In light of the fact that a significant number of languages in existence lack sufficient annotated data, knowledge transfer from high-resource languages to low-resource languages can serve as a viable solution to compensate for the paucity of data [491,492]. Developing robust multi-lingual NER systems that can perform across multiple languages will achieve more comprehensive knowledge graphs, linking entities from different languages. It is valuable because it may lead to a united concept representation system covering different languages. On the other hand, the task of developing multi-lingual NER systems is fraught with difficulties, primarily due to the inherent dissimilarities in entity types and language structures across different languages. As a result, aligning entities and transferring knowledge learned from one language to another can present significant challenges for multi-lingual NER systems.

Unified framework for NER. In the real-world scenario, there exist flat-named entities, nested entities, and discontinuous entities. Most NER-related studies only focus on the combination of flat with nested entities or flat discontinuous entities. Both of them cannot recognize all kinds of entities. Developing a unified framework to simultaneously handle such a problem becomes an urgent need for NER [430]. Hierarchical concept representation knowledge bases may provide a preliminary ontology that can be used for organizing entities and their relationships. However, most of the ontology systems were manually developed by experts. This manually constructed knowledge may be invalid in specific application scenarios. A potential avenue for future research in NER is the development of a unified and robust framework for organizing entities. Such a framework could facilitate the creation of comprehensive knowledge graphs that capture the relationships between entities and can better support downstream tasks.

Continual-learning for NER. Humans exhibit a remarkable aptitude for transferring acquired knowledge from one task to another and retain their ability to perform the former task even after learning the latter. This ability is called continuous learning or life-long learning, which a regarded as an important characteristic of an intelligent system. Also, such ability can help us continue to use already deployed models when a new class of entity to be identified appears, rather than developed a new model from scratch [493]. There are some exploratory studies started to pay attention to such a problem. However, a satisfactory solution has not been found yet and existing methods still suffer the severe Catastrophic Forgetting [494–496]. Continual learning is a critical skill for NER because NER is corpus-dependent. It is very important to update entity collections and the associated label sets, when a new corpus arrives [497]. In this case, detecting new entities and new labels with a former trained NER model represents a challenging yet highly promising research avenue.

5. Concept extraction

Concept extraction is a process to extract concepts of interest from the text. To our best knowledge, the task of computational concept extraction was first proposed by Montgomery [498], which analyzed the next 5 years of evolutionary progress in contemporary military message routing systems, with a focus on their transition towards becoming more advanced and knowledge-based systems. They argue that taxonomic hierarchies could be constructed to allow property inheritance of concepts, and therefore to perform rudimentary inference and analogic reasoning based on the taxonomies. Montgomery [498] also highlighted two important sub-tasks of concept extraction for the next-generation knowledge-based systems from the perspective of 1982, namely lexicon development and conceptual structure construction.

Recent research on concept extraction has been conducted in various fields of AI research, including natural language processing (NLP) and data mining [499]. Keyphrase generation [500] is one of the most common concept extraction tasks. It is a summarization task focusing on extracting keyphrases from a full passage to help readers quickly understand the passage, where keyphrases can be understood as the important concepts within a passage. Methods for keyphrase extraction can be both extractive (copying from existing words) and abstractive (not copying but summarizing and abstracting from existing texts). The process of generating keyphrases facilitates the creation of a lexicon that corresponds to a specific set of concepts. Another stream of concept extraction aims at the development of ontological knowledge bases to represent, e.g., commonsense knowledge [501], hypernym and synonym knowledge [502], sentic knowledge [503]. These tasks tried to extract concepts to fit into pre-defined knowledge structures. Then, the structured knowledge can be directly used in downstream tasks.

Current concept extraction research is also grounded on related application scenarios, such as clinical concept extraction [14], course concept extraction [504], and patent concept extraction [505]. Clinical concept extraction is to transform massive unstructured electronic health records data into structured data; Course concept extraction is to extract important phrases in course captions to help to understand. Among them, clinical concept extraction is very similar to the information extraction task in NLP which aims at extracting most of the details in the unstructured text. Course and patent concept extraction are more similar to summarization tasks in NLP that target extracting important phrases.

The main difference between concept extraction and NER tasks is that the extracted concepts or keyphrases are not identified by predefined entity classes. In contrast, they reflect the general idea of their contexts or target domain whose concepts are being discussed, while the goal of NER is to extract important factual information from the text. However, there are overlaps between NER and concept extraction when some concepts of interest, e.g., proper nouns can be also defined as named entities. Many domain-specific concept extraction tasks, e.g., clinical concept extraction, course concept extraction, and patent concept extraction can also be categorized as NER tasks because they aim at extracting concepts that are related to specific events. These events are also factual information. We review them in this section because they define themselves as concept extraction tasks in their original works. It also has become a trend of domain-specific concept extraction.

Another related field is relation extraction, which is a sub-field of information extraction. Relation extraction extracts information from raw text and represents it in the form of a semantic relation between entities [506]. The main difference is that, relation extraction targets at extracting relations between entities, while concept extraction targets at extracting noun entities. In knowledge graph development, relation extraction can help to connect nodes of concepts with purposeful relationships.

Concept extraction has also accelerated and contributed to multiple downstream applications, such as sentiment computing [503], information retrieval [507], commonsense explanation generation [508]. These applications mostly leverage explicitly extracted concepts.

Previous survey on concept extraction on focuses on clinical concept extraction [14], which is a particular application field of concept extraction. In this section, we provide a more comprehensive review on concept extraction.

5.1. Theoretical research

5.1.1. Exemplar theory

Medin and Schaffer [509] argued that concepts are represented by a collection of particular exemplars or individual instances that are linked to the category. When we categorize an instance, we compare it with multiple specific exemplars of the category. This is different from Prototype Theory where a new instance is categorized by comparing the instance to the abstract prototype of the category (see Section 4.1.1). Medin and Schaffer [509] formed the task of concept categorization as a classification task, and conducted experiments with 32 participants. The experiments showed that the classification judgments made by participants were impacted by various factors. These factors included the extent of resemblance between the probe item and exemplars previously acquired, the number of prior exemplars that shared resemblances with the probe item, and the similarity present both within and between the categories of the previously learned exemplars. For concept extraction and categorization, Exemplar Theory may suggest that models may take categorized instances into account when categorizing a new instance.

5.1.2. Semantic primitives

Wierzbicka [510] believed that it is possible to describe every human language by using a limited number of universal semantic primitives. These primitives are representative of fundamental concepts that form the basis of human communication and thinking. Wierzbicka [510] established 64 universal semantic primes, which consist of basic words or ideas that cannot be defined in relation to more elementary concepts. However, these primes can be utilized to describe all other concepts present within a language. Semantic Primitives suggest that concepts should be organized as multiple layers from the concrete to abstract ones. Decision-making that runs on concrete concepts can be completed through the upper-level abstract concepts that contain those concrete concepts. Thus, it is critical to represent the hierarchical and linking relationships between concepts. There are other theories mentioned before, e.g., Frame Semantics (see Section 2.1.4), that may guide concept structure development. Frame Semantics highlights the connection of related concepts, while Semantic Primitives suggest the hierarchical relationships between concepts and the distinction between primitive concepts and others.

5.1.3. Conceptual spaces

Gardenfors [511] defined concept as the “theoretical entities that can be used to explain and predict various empirical phenomena concerning concept formation”. The author believed that concept representations are multi-dimensional, where each dimension is indicative of a different characteristic or property associated with the concept. For example, one could represent the concept of a car within a conceptual space that includes dimensions such as size, speed, color, and shape. This is very similar to current vectorial representations of words or entities in NLP, while the dimensionality of Conceptual Spaces is explainable by concept properties. Gardenfors [511] also placed significant emphasis on the role of context in understanding and representing concepts. This is due to the fact that different contexts may emphasize different features or dimensions of concepts. Then, the connections between concepts are determined by the relationships between their property similarity in the conceptual space. For example, “dog” and “cat” are similar in the animal concept space, because their properties are similar; “mammals” can be separated from “reptiles” by a property difference boundary, although both are in the animal space. This may encourage concept extraction tasks to extract both concept entities and properties associated in contexts. This is because properties define how concepts are connected from the view of Gardenfors [511].

5.2. Annotation schemes

From the goal of the keyphrase annotation aspect, there are in general two types of annotation schemes for keyphrase extraction-liked concept extraction. The first is to precisely select existing keyphrases from input text, but not to create semantically-equivalent phrases. The second is to both select existing keyphrases and create “absent keyphrases” that are necessary but do not exist in the input text [512].

From the format of assigned annotations aspect, there are in general two annotation schemes as well. The first scheme is to directly give the keyphrases existing in the source text. The second scheme treats the keyphrase extraction task as a sequence labeling task, and assigns a label to each of the tokens in source text [512]. The assigned labels in the current dataset follow a BIO scheme defined in Table 10. Specifically, three labels are used: B (Beginning), I (Inner), and O (Other).

5.3. Datasets

The surveyed popular concept extraction datasets and their statistics can be viewed in Table 15. Overall the main thread of dataset development is (1) larger scale of datasets; (2) attending to both extractive keyphrases and abstractive keyphrases; (3) more fine-grained annotations for tags; (4) more application domains. Hulth [512] proposed one

Table 15
Concept extraction datasets and statistics. KE denotes Keyphrase Extraction. ClCE denotes Clinical Concept Extraction. CoCE denotes Course Concept Extraction. PCE denotes Patent Concept Extraction.

Dataset	Task	Source	# Samples	Reference
Inspec	KE	Inspec database	2,000	Hulth [512]
NUS	KE	Google SOAP API	211	Nguyen and Kan [513]
Krapivin	KE	ACM Digital Library	2,304	Krapivin et al. [514]
SemEval2010	KE	ACM Digital Library	244	Kim et al. [515]
Twitter	KE	Twitter	1,000	Zhang et al. [516]
KP-20K	KE	ACM Digital Library, ScienceDirect, and Web of Science	567,830	Meng et al. [517]
CCF	KE	China Computer Federation	13,449	Wang et al. [518]
MLDBMD	KE	Academic Conferences	128.1k	Li et al. [519]
TempEval	ClCE	Mayo Clinic	600	Bethard et al. [520]
i2b2-2010	ClCE	Clinical Records	826	Uzuner et al. [521]
n2c2-2018	ClCE	Clinical Records	505	Henry et al. [522]
MIMIC	ClCE	MIMIC-III Database	1,610	Gehrmann et al. [523]
MOOCs	CoCE	Coursera and XuetangX	4375 videos	Pan et al. [524]
EMRCM	CoCE	Chinese Textbooks	3,730 pages	Huang et al. [525]
USPTO	PCE	USPTO Database	94,000	Liu et al. [505]

of the first keyphrase extraction datasets, termed Inspec. Their dataset is based on the scientific papers under *Computers and Control*, and *Information Technology* disciplines in the Inspec database. The keywords used in the scientific papers are selected as the keyphrases. Abstracts are used as the keyphrase extraction context. Keywords in scientific papers are used as keyphrases. Each abstract has two sets of keywords: a set of controlled terms, i.e., terms restricted to the Inspec thesaurus; and a set of uncontrolled terms that can be any suitable terms that may or may not be present in the abstracts. They collected 1000 abstracts as a train set, 500 as a validation set, and 500 as a test set.

```
abstract: "[ ‘A’, ‘scalable’, ‘model’, ‘of’, ‘cerebellar’,
‘adaptive’, ‘timing’, ‘and’, ‘sequencing’, ‘:’, ...]"
doc bio tags: "[ ‘O’, ‘B’, ‘I’, ‘O’, ‘B’, ‘I’, ‘I’, ‘O’, ‘O’, ‘O’,
...]"
extractive keyphrases: "[ ‘scalable model’, ‘cerebellar adaptive
timing’, ... ]"
abstractive keyphrase: "[ ‘cerebellar sequencing’, ...]"
```

Nguyen and Kan [513] proposed the NUS dataset with the motivation that keyphrase extraction requires multiple judgments and cannot rely merely on the single set of author-provided keyphrases. They first used Google Search API to retrieve scientific publications, and then recruited student volunteers to participate in manual keyphrase assignments. They finally collect 211 documents, each with two sets of keyphrases: one is given by the original authors of the paper, and the other is given by student volunteers. The data format of NUS is the same as Inspec [512]. Krapivin et al. [514] proposed the Krapivin dataset, consisting of around 2000 scientific papers as well as their keywords assigned by the original authors. The scientific papers were published by ACM in the period from 2003 to 2005, and were written in English. One of the novelties of this dataset is that the text data in the scientific papers were collected with three distinct categories: title, abstract, and main body. They finally collect 460 test data and 1.84k validation data. The data format is similar to Inspec [512] but has a title and body in addition to the abstract.

SemEval-2010 Task 5 [515] is on automatic keyphrase extraction from scientific articles. Input for this task is a document from either of the four domains: distributed systems, information search, and retrieval, distributed artificial intelligence, and social and behavioral sciences. Outputs are manually annotated keyphrases for the document. This dataset contains 144 documents as a train set, and 100 documents as a test set. It also selects 40 documents from the train set to compose a trial set. For each set, documents are evenly distributed from the four topics. The annotation follows the first scheme in Section 5.2. The data format is the same as Inspec [512]. Zhang et al. [516] constructed a keyphrase extract dataset from Twitter using an automatic text mining method. Their core assumption is that hashtags in a tweet can be used as keyphrases for the tweet. To construct the dataset, they first collected 41 million tweets, and then filtered them which contain non-Latin tokens. URL links, and reply tweets were removed. Thus, the remaining text only contains tweets and a hashtag. They finally kept 110K tweets. To evaluate the quality of the collected tweets, they sampled 1000 tweets and chose three volunteers to score them. As a result, 90.2% tweets are suitable, and 66.1% are perfectly suitable. The annotation follows the first scheme in Section 5.2.

```
tweet: "Hard to believe it but these are REAL state alternatives
to taking Obamacare
funds from the gov’t (via @Upworthy)"
keyphrase: "obamacare"
```

Pan et al. [524] proposed a keyphrase extraction dataset, where data were sourced from online course captions. Labels are existing phrases in the captions. The courses are computer science and economics courses, selected from two famous MOOC platforms — Coursera and XuetangX. Labels were first filtered from captions using automatic methods and then annotated by two human annotators. A candidate concept was only labeled as a course concept if the two annotators were in agreement. As a result, they collected captions from 4375 videos, and 16720 labeled concepts.

```
course caption: "You might learn how to write a bubble sort and
learn why a bubble sort is not as good as a heapsort."
keyphrase: "[ ‘bubble sort’, ‘heapsort’ ]"
```

KP-20K [517] is a testing dataset, where the input texts are titles and abstracts of computer science research papers collected from ACM Digital Library. The labeled keyphrases are the keyphrases shown in the research papers. The annotation follows the second scheme in Section 5.2, since the keyphrases given by authors were not necessarily existing keyphrases in the papers. KP-20K has the same data format as Inspec.

Huang et al. [525] were motivated to automatically construct an educational concept map. The educational concept map shows concepts that will be learned in courses, as well as the temporal relation between the concepts (e.g., to learn concept A, it is a prerequisite to learn concept B; Concept A and concept B can help with the understanding of each other). To construct the dataset written in Chinese, they first used OCR to obtain the text from textbooks, then manually labeled key concepts for each textbook (as “key concept” or “not key concept”) and finally manually annotated the relationships among the labeled key concepts (as “w_i is w_j’s prerequisite”, “w_i and w_j has collaboration relationship”, or “no relationship”). As a result, they collected 3730 pages in textbooks, 1092 key concepts, 818 prerequisite relations, and 916 collaboration relations. However, in their GitHub repo, only keyphrases and relations between keyphrases can be found, while the text cannot be found.

```
keyword: "[ ‘average’, ‘weighted average’, ... ]"
relation: "[ ‘average : weighted average’, ... ]"
```

There are concept extraction datasets focused on a specific domain, e.g., clinical concepts (TempEval [520], i2b2-2010 [521], n2c2-2018 [522], and MIMIC [523]), course concepts (MOOCs [524], and EMRCM [525]), and patent concepts (USPTO [505]). They also followed keyphrase extraction setups, whereas the targets are to extract concepts of interest.

5.4. Knowledge bases

Besides classical lexicon resources such as WordNet, encyclopedias (including Baidu Encyclopedias and Wikipedia) can also be used to provide external knowledge for concepts [524]. Methods for extracting concepts based on embedding techniques may encounter issues with low frequency, where some of the concepts have infrequent occurrences. Pan et al. [524] utilize word embeddings [23], which is trained on encyclopedias, to obtain the semantic embedding for each concept. Inspec database is a scientific and technical database storing scientific papers. The papers of this database have been used to construct a keyphrase extraction dataset (see Table 16).

5.5. Evaluation metrics

The field of concept extraction also uses Precision, Recall, and F1-score as evaluation metrics. Some keyphrase extraction research considered the task as an information retrieval task. Then, the information retrieval metric, e.g., mean average precision (MAP) was also used for keyphrase extraction as the main measure. It is calculated by taking the average of the average precision scores for each query in a dataset.

$$MAP = \frac{1}{n}\sum_{i=1}^{n} Avg\_Precision_i, \tag{9}$$

where n is the total number of queries. $Avg\_Precision_i$ denotes the averaged precision of query i. In the context of keyphrase extraction, the MAP score is determined by comparing the generated list of keyphrases with a predefined gold standard set, and evaluating the average precision of the top n keyphrases, where n corresponds to the total number of keyphrases in the gold standard set. Each generated keyphrase is considered as a query; The gold standard set serves as the relevant document.

5.6. Annotation tools

Since the annotation schemes of concept extraction are similar to that of NER. The aforementioned NER annotation tools can also be used for annotating concept extraction data. Numerous studies have investigated the utilization of pre-existing keywords in scientific publications [512–515,526] or hashtags in tweets [516], whereby such in-context information can serve as labels without requiring additional annotation efforts, provided that the labels align with the research objectives.

Table 16
Useful knowledge bases for concept extraction.

Name	Knowledge	#Entities	Structure
WordNet	Lexical	155,327	Tree
Baidu Encyclopedia	World	6,223,649	Unstructured
Wikipedia	World	9,834,664	Unstructured
Inspec	Science	20,000,000	Unstructured

5.7. Methods

5.7.1. Keyphrase extraction

The task of keyphrase extraction is to obtain keyphrases from a document to represent and summarize the document with the keyphrases. There are generally two trends of methods, namely extractive keyphrase extraction and generative keyphrase extraction.

Extractive methods appear first but have a systematic disadvantage in that they can only extract existing phrases in the documents. For example, [517] argued that in addition to present keyphrases, there are also absent keyphrases, which can better summarize a document but do not explicitly present in the document. Generative methods, however, can generate every possible word. Therefore generative methods can alleviate the disadvantage of extractive methods, but might be more difficult because it requires a model to accurately catch the “semantic meaning” of a document to precisely generate a keyphrase.

A. Extractive Keyphrase Extraction

Zhang et al. [516] focused on the task of keyphrase generation on Twitter data, and framed this task as a sequence labeling task. They proposed a joint-layer RNN model. For each input token, the joint-layer RNN model outputs two indicators ($\hat{y}_1$ and $\hat{y}_2$), where $\hat{y}_1$ has two values $True$ and $False$, indicating whether the current word is a keyword. $\hat{y}_2$ has 5 values $Single$, $Begin$, $Middle$, End and Not indicating the current word is a single keyword, the beginning of a keyphrase, the middle of a keyphrase, the ending of a keyphrase, or not a part of a keyphrase, respectively. Their experiments show that the joint-layer RNN model outperforms both the vanilla RNN model and the LSTM model. However, when $\hat{y}_1$ and $\hat{y}_2$ have contradictions, it might be hard to find an optimal strategy to determine which indicator to refer to. In addition, joint-layer RNN can only extract an existing sequence as a keyphrase, but cannot abstractively obtain a (non-existing but better) keyphrase.

Wang et al. [518] hypothesized that the performance of keyphrase extraction could be improved in the unlabeled or insufficiently labeled target domain by transferring knowledge from a resource-rich domain. They accordingly proposed a topic-based adversarial neural network (called TANN) that can learn transferable knowledge across domains efficiently by performing adversarial training. The experiment section shows that TANN largely outperforms joint-layer RNN [516].

Li et al. [519] proposed an unsupervised method for concept mining, which was motivated by the fact that supervised methods might be hard to generalize to unseen domains. They assumed that the quality of an extracted concept can be measured by its occurrence contexts and proposed a pipeline method for concept mining. The method first populates many raw concepts extracted from text, and then evaluates the concepts by comparing the embedding of concepts against the current local context.

Al-Zaidy et al. [527] identified two limitations of previous supervised approaches: (1) They classify the labels of each candidate phrase independently without considering potential dependencies between candidate phrases. (2) They do not incorporate hidden semantics in the input text. Correspondingly, [527] addressed keyphrase extraction as a sequence labeling task, and proposed a model named Bi-LSTM-CRF that unite both the advantages of LSTM (capturing semantics) and CRF (Conditional Random Field, capturing dependencies,[528]). Their results show that Bi-LSTM-CRF outperforms CopyRNN [517] by a large margin.

Fang et al. [529] hypothesized that previous extractive methods ignore structured information in the raw textual data (title, topic, and clue words), which might lead to worse performance. They accordingly proposed a model named GACEN that can utilize the title, topic, and clue words as additional supervision to provide guidance. GACEN also utilized CRF to model dependencies in the output. The experiment section shows that GACEN outperforms Joint-layer-RNN [516] and CopyRNN [517].

B. Generative Keyphrase Extraction

Meng et al. [517] were motivated that classic keyphrase generation methods can only extract the keyphrases that appear in the source text. Those methods are unable to reveal and leverage the full semantics for keyphrase ranking. Consequently, they proposed an RNN-based generative model incorporating a copying mechanism [530] (named with CopyRNN), which can generate absent keyphrases. Their method uses an encoder–decoder architecture to catch the semantics of the input text.

Previous methods such as [517] suffered from both coverage (not all keyphrases are extracted) and repetition (similar keyphrases are extracted) problems. For the coverage issue, [526] integrated a coverage mechanism [531] into their approach, which enhances the attention distributions of multiple keyphrases in order to cover a wider range of information within the source document and effectively summarize it into keyphrases. For the repetition issue, they constructed a target side review context set that contains contextual information of generated phrases.

Ye and Wang [532] believed that although sequence-to-sequence (seq2seq) models have achieved good performance, model training still relies on large amounts of labeled data. Correspondingly, they leveraged unsupervised learning methods such as TF-IDF and self-learning algorithms to create keyphrase labels for large amounts of unlabeled data. Then, they train their model with a mixture of self-labeled and labeled data together for training. They also used multi-task learning to train their model. Experiments show that their performance outperforms previous works.

Chen et al. [533] argued that prior research on keyphrase generation has treated the document title and main body in the same manner, overlooking the significant role that the title plays in shaping the overall document. They accordingly proposed a Title-Guided Network (TG-Net) where the title is additionally employed as a query-like input to particularly assign attention to the title. The performance of TG-Net outperforms CopyRNN [517]. Their ablation study also shows the importance of additional attention to the title.

5.7.2. Structured concept extraction

Compared with keyphrase extraction-liked concept extraction, structured concept extraction aimed to develop an ontology where concepts are connected with each other by certain relationships. Here, we introduce three knowledge bases resulting from concept extraction: WordNet, ConceptNet, and SenticNet. Out of them, WordNet focuses more on a word-level ontology, ConceptNet focuses more on a concept-level ontology (e.g., also including phrases for concepts), and SenticNet is a concept-level ontology focusing on contributing to sentiment analysis tasks.

WordNet is a manually developed knowledge base, where words and concepts are hierarchically organized. Snow et al. [502] proposed a taxonomy induction method to expand WordNet 2.1 concepts by automatic noun hyponym acquisition, achieving 10,000 novel synsets with 84% precision. Compared to previous methods that relied on individual classifiers to uncover new relationships based on pre-designed or automatically extracted textual patterns, the proposed approach considers input from multiple classifiers to enhance the overall structure of the taxonomy and prioritizes the optimization of the entire taxonomy structure with a probabilistic architecture. Snow et al. [502] also proposed an (m,n)-cousin classification-based model to learn coordinate terms, which allows it to integrate heterogeneous evidence from different classifiers and choose the correct word sense to which to attach a new hypernym. The evaluation of the inferred taxonomies produced by the algorithm was conducted by directly comparing them with the WordNet 2.1 taxonomy. This was achieved by testing each taxonomy using a set of human judgments of noun pairs sampled from newswire text, to determine the hypernym and non-hypernym relationships.

ConceptNet [501] grew out of Open Mind Common Sense project that aimed at commonsense acquisition. Contributors delivered knowledge by fulfilling blanks within a sentence, For example, given "[ ] can be used to [ ]", the concepts, e.g., "ink" and "print" and the associated relationship "UsedFor" can be obtained. ConceptNet aimed to obtain and structure concepts automatically from natural language. It obtained concepts (the nodes) in the form of noun phrases, verb phrases, adjective phrases, prepositional phrases, or complete verb phrases [501]. The edges of ConcepNet are predicates that represent the relationships between two concept nodes, such as "IsA", "PartOf", UsedFor, and more. Havasi and Speer [501] defined 21 basic relation types. In the latest ConceptNet 5.5 [534], the relations are increased to 36. Concepts and predicates were obtained via pattern matching. Each collected sentence is compared with pre-defined regular expressions, e.g., "NP is used for VP"(UsedFor), "NP is a kind of NP"(IsA), "NP can VP" (CapableOf). NP (noun phrases) and VP (verb phrases) are concepts, while "UsedFor", "IsA", and "CapableOf" are predicates. In the case of a complex sentence that contains several clauses, the patterns are employed to extract a simpler sentence from it, which can then be subjected to the concept and predicate extraction process. To evaluate ConceptNet, its assertions were compared with those in similar lexical resources to determine their alignment.

SenticNet is a commonsense knowledge base that is used for affective computing. The concepts were extracted by a graph-based semantic parsing method [535] and assigned with sentiment polarity labels. Sentences are divided into chunks, e.g., "go walk", first. Then, verb-noun chunks are normalized by stemming, and included in the concept set. The PoS-based bigram algorithm is used to extract object concepts. To capture event concepts, the approach explores matches between object concepts and normalized verb-noun chunks. Finally, single-word concepts, e.g., "house" that have appeared in the clause as multi-word concepts "beautiful house" are deemed redundant and are therefore excluded. In the following version of SenticNet [536], the authors proposed an automatic method to discover primitives from the SenticNet concepts, based on hierarchical clustering and dimensionality reduction. Thus, the "animal" concept can be identified as the primitive of "cat", "dog", or "pet". Later, [537] proposed a pipeline method for concept extraction, which is used for expanding SenticNet with multi-word expressions. They first deconstructed text using sentence chunking, semantic parser, and PoS tagging. Then, verb and noun chunks are extracted and normalized as concepts. The proposed method offers novel contributions in utilizing morphology for syntactic normalization and employing primitives for semantic normalization. The method was evaluated on a sentiment analysis task, achieving explainable and primitive- and concept-level sentiment analysis via algebra operations. The latest version of SenticNet [503] offers the function that sentiment predictions can be effectively conducted on the primitive level, mitigating symbol grounding problems.

An important task of concept extraction is to abstract concept representations from entities. Unlike SenticNet which obtains abstract concepts (primitives) by selecting the most typical entities from a group of extracted similar entities [536], [538] proposed a conceptualization method that can directly abstract concepts from input text. The task is realized in the metaphor identification and interpretation domain. The authors aimed to generate concept mappings from metaphorical word pairs to explain the metaphoricity of the word pairs. For example, given "*blind* alley", "STREET is ADULT" can be automatically generated. This work is the realization of conceptual metaphor theory [539] (see Section 6.1.3) that the generated concept mapping explains the mapping of source (e.g., ADULT) and target (e.g., STREET) concepts of a metaphor. The conceptualization (e.g., from "alley" to "STREET") was achieved by selecting the most appropriate hypernym on the chain from the leaf node of "alley" to the root node "entity" in WordNet. The most appropriate hypernym is defined as the node that can cover the major senses of the leaf, meantime, keeping it as concrete as possible.[33] The conceptualization and concept mapping method was evaluated on a metaphor identification task, yielding better performance and explainability on the task. Subsequently, within MetaPro Online [540], the conceptualization algorithm is synergistically integrated with sequential metaphor identification and interpretation techniques, culminating in the attainment of end-to-end concept mapping generation from full sentences.

5.7.3. *Domain-specific concept extraction*

A. Clinical Concept Extraction

The task of clinical concept extraction is to extract structural information from unstructured clinical narratives [14]. Li and Huang [541] constructed a dataset for a seminal task called "UTA-DLNLP at SemEval-2016 Task 12" for clinical concept extraction. A system developed for this task should task raw clinical notes or pathology reports as input, and identify event expressions consisting of the "the spans of the expression in the raw text", "contextual modality", "degree", "polarity", and "type". As a baseline for this task, they propose a convolutional neural network to learn hidden feature representations for predictions, taking text and part-of-speech tags as input.

Liu et al. [542] adopted BiLSTM to recognize the entity in clinical text. They found that BiLSTM outperforms the CRF baselines. Gehrmann et al. [523] compared CNN with classic rule-based methods, bag of words, n-grams, and embedding-based logistic regression. They found that CNN is a valid alternative to rule-based and classic NLP methods, and should be further investigated. Yang et al. [543] comprehensively explored 4 widely used transformer-based architectures, including BERT [8], RoBERTa [9], ALBERT [544], and ELECTRA [545]. They compared the 4 models to long short-term memory conditional random fields (LSTM-CRFs) [546] baselines and found that transformer-based models are effective for clinical concept extraction tasks. Lange et al. [547] proposed a joint model for both clinical concept extraction and de-identification tasks. De-identification is important since in some clinical concept extraction scenarios, the privacy of patients should be protected. They hypothesized that jointly modeling the two tasks can be beneficial, and proposed two end-to-end models. One is a multitask model where the tasks share the input representation across tasks; the other is a stacked model, which used the privacy token predictions to mask the corresponding embeddings in the input layer and only use the masked embeddings for concept extraction. They found that the performance of the concept extraction model can be improved by training and evaluating it on anonymized data, thereby confirming their initial hypothesis.

B. Course Concept Extraction

In tasks involving the extraction of course concepts, the concepts are typically defined as the knowledge concepts that are taught in the course videos, as well as the related topics that aid in the students' comprehension of the course videos [524]. Identifying course concepts at a fine level is very important, as students with different backgrounds need different concepts to quickly understand the main content of a course [524].

Pan et al. [524] contributed the first attempt to systematically investigate the problem of course concept extraction in MOOCs. in the past, course concepts were presented by instructors at a general level, with only a few concepts being covered in an entire course video. However, they emphasized the significance of identifying course concepts at a granular level, i.e., automatically identifying all course concepts from each video clip, to facilitate easier comprehension. They identified a challenge for the task that the course concept appears at a low frequency mainly because the different courses have different concepts. They accordingly proposed to utilize word embedding to catch the semantic relations between words and incorporate online encyclopedias to learn the latent representations for candidate course concepts. They also proposed a graph-based propagation algorithm to rank the candidates based on learned representations.

Wang et al. [548] argued that external knowledge must be involved to solve the concept extraction problem and proposed to utilize both the structured and unstructured data in Wikipedia to provide external knowledge to concept extraction. Their results show that their method outperforms prior works [524].

C. Patent Concept Extraction

Liu et al. [505] developed a framework to extract technical concepts from patents. Patent documents have different structures than other documents. For instance, they have "title", "abstract", and "claim", which exhibit a multi-level of information. Motivated by this, the authors proposed a framework named UMTPE, which can effectively leverage multi-level information to extract concepts.

5.8. *Downstream applications*

5.8.1. *Sentiment computing*

SenticNet 7 [503] is a neuro-symbolic sentiment analysis system, based on SenticNet knowledge base. It assumes that concepts that share the same primitive would have similar sentiments. One can use algebra operations to achieve sentiment analysis with the symbolic and structural knowledge base. Incorporating a symbolic knowledge base and a transparent algorithm provides SenticNet's reasoning process with the benefit of interpretability and accuracy.

Li et al. [549] proposed a neuro-symbolic system for conversational emotion recognition. ConceptNet was used as a knowledge base to acquire commonsense knowledge out of context. For example, if a person mentions that he will "chop all onions we have and cry", another conversation participant expresses "disgust" emotion. This is because "onion IsA lacrimator" is a commonsense in ConceptNet. Such a commonsense cannot be obtained from the dictionary meanings of "onion" and the context, while ConceptNet commonsense knowledge provides the evidence and explainability to infer such an emotional status from the context. The authors used an utterance dependency parser and a neural network to learn symbolic knowledge to enhance the explainability and accuracy of their method.

By using the concept mapping method from the work of Ge et al. [538], Han et al. [550] used concept mappings to support depression detection and explanation. The hypothesis is that depression patients may have similar cognition patterns that are reflected in their metaphorical expressions. Thus, they used concept mappings as additional features besides tweets. The concept mappings were generated from tweets that contained metaphors. They also proposed an explainable encoder that can identify significant concept mappings that contribute to depression detection. The concept mappings also improve the accuracy of depression detection, besides explaining the common concept mapping patterns.

5.8.2. *Information retrieval*

Xiong et al. [507] manually analyzed the potential problems of a literature search website SemanticScholar.org, and found that the issue of "Concept Not Understood" represents one of the most significant challenges. The reason is that previous methods measure similarity based on text, but not on their semantic embeddings. As a result, they proposed an embedding-based similarity matching method, which extracts the concepts in both query and documents and measures the similarity between these concepts to obtain the similarity between a

[33] Intuitively, "entity" can cover all possible senses of the "alley" in WordNet, while it is not the ideal concept representation of "alley", because it is too abstract. Thus, the authors aimed at a concrete concept representation that can cover the majority senses of a word.

query and a document. Liu et al. [551] used extracted knowledge concepts as one of the inputs to obtain a unified semantic representation for educational excises. The representation is further used to retrieve similar excises based on similarity with other representations.

5.8.3. Dialogue systems

Young et al. [552] integrated commonsense knowledge from ConceptNet in their dialogue system. They believed that in human dialogues, individuals responding to each other is not dependent on the most recent utterance only, but also on recollecting pertinent information related to the concepts addressed within the dialogue, e.g., commonsense. Thus, in retrieval-based dialogue generation, the model considers both the message content and relevant commonsense knowledge to effectively choose a suitable response.

Huang et al. [553] proposed a new dialogue coherence evaluation matric, termed Graph-enhanced Representations for Automatic Dialogue Evaluation (GRADE). Liu et al. [554] argued that traditional BLEU-liked statistic-based metrics are biased in response coherence. Thus, [553] were motivated to propose a metric that measures the coherence by the topics of utterances. They believe that a cohesive exchange of dialogues is characterized by a seamless transition between topics. Thus, they used a ConceptNet-based method to construct topic-level dialogue graphs. The topic-level dialogue graphs were constructed by connecting the concepts that are extracted from utterances. The edge was weighted and undirected, which was derived from the shortest path between two nodes in the ConceptNet. Such an evaluation metric can better represent the coherence of topics between utterances because it measures the relatedness of concepts from different utterances.

5.8.4. Commonsense explanation generation

Fang and Zhang [508] grounded concept extraction in the context of commonsense explanation generation. Commonsense explanation generation aims to generate an explanation in natural language to explain the reason why a statement is anti-commonsense. For example, given “he took a nap in the sink”, the model aimed to generate “a sink is too small and dirty to take a nap in”. The concepts, “small” and “dirty” (bridge concepts), are obtained via a prompt-tuning method. The authors developed a masked word prediction template to query the bridge concepts that are most likely to appear in the “mask” position. Then, they use a generator to generate the explanation with the concatenation of the original statement and the discrete bridge concepts. This method improves the explainability in explaining why a statement is anti-commonsense.

5.9. Summary

A concept is an abstract idea that is reflected in the mind. Concept extraction is the foundation of detecting the main idea of a context and developing conceptual knowledge bases. Related theoretical research showed that concepts may be abstracted from multiple specific exemplars [509] or prototypes [384]. There are limited primitives that construct human cognition and reasoning, which are the foundation of complex concepts [510]. According to [511], conceptual space is multi-dimensional. The similarity between concepts can be measured by the similarity between concept properties. These theoretical research works frame the tasks of concept extraction from the perspectives of lexicon development and conceptual structure construction. On the other hand, current computational concept extraction methods divide this task into three categories, namely keyphrase extraction, structured concept extraction, and domain-specific concept extraction. We found that the existing computational approaches inadequately address the tasks that have been put forth by the academic community’s theoretical research. Although current concept extraction methods are limited, this task has greatly improved the explainability of downstream tasks such as sentiment computing, information extraction, and counter-commonsense recognition.

5.9.1. Technical trends

Within the domain of keyphrase extraction, generative keyphrase extraction takes advantage of generating “absent keyphrases”, compared to extractive keyphrase extraction. Both tasks followed the general development of the NLP fields. They likely considered the task as a sequence labeling task (extractive keyphrase extraction) or a generation task (generative keyphrase extraction), and used typical NLP frameworks, e.g., sequence labeling and sequence to sequence frameworks. However, it is unclear if these general NLP frameworks have really learned how to summarize the main idea of context or just have learned by label distributions. There were no task-specific mechanisms proposed to explicitly learn the keyphrase extraction task on the concept level, with an explainable decision-making process. On the other hand, keyphrase extraction-based concept extraction is helpful for obtaining concept lexicons. However, compared to structured concept extraction, keyphrase extraction cannot learn the relationships between concepts. The theoretical research of Conceptual Spaces from [511] suggested that the similarity between concepts can be measured by their properties. It suggests that keyphrase extraction-based concept extraction should consider extracting properties together with keyphrases. Thus, the later works can use keyphrases and the associated properties to structure concepts by similarities (see Table 17).

In contrast, structured concept extraction research likely utilized statistical learning and syntactic parsing methods. This is because the aim of structured concept extraction is to develop a large knowledge base or detect structured relationships between concepts. Labeled data are insufficient in these areas. Thus, unsupervised methods are preferred. However, the concept knowledge base development is task-specific. As a result, the concepts in different knowledge bases share different relationships. For example, ConceptNet aimed to parse concepts sharing 36 commonsense relationships; Stanford WordNet [502] was expended in synonyms and hypernyms relationships; SenticNet grouped concepts and extract primitives for sentiment computing; [538] abstracted concepts for concept mappings. Then, the evaluation of different concept extraction methods is different. Most of the evaluation was implemented on different downstream tasks. It shows

Domain-specific concept extraction is very similar to NER tasks. They used graph, machine learning methods, and external knowledge, e.g., encyclopedias and Wikipedia to discover concepts in a domain, e.g., clinical, course, or patent concepts. Similar to keyphrase-based concept extraction, these domain-specific concept extraction methods did not try to structure concepts after extraction. This is important because it distinguishes concept extraction from current NER tasks in specific domains.

5.9.2. Application trends

Concept extraction methods and their product, e.g., knowledge bases have been widely used in downstream tasks, e.g., sentiment computing, information retrieval, dialogue systems, and commonsense explanation generation. Compared to other low-level semantic processing techniques, the roles of concept extraction are more diverse in downstream applications. For all the surveyed downstream tasks, the products of concept extraction can be used as additional features to improve model performance on downstream tasks. On the other hand, concept extraction techniques can be used as a parser to obtain knowledge from unstructured text. The structured concepts with certain relationships can also improve the explainability of a downstream task model, e.g., explaining anti-commonsense [508] and concept mapping patterns of depressive patients [550] (see Table 18).

In the era of PLM and LLM, it seems many complex tasks can be achieved from end-to-end with deep neural networks. However, black box-liked neural networks prevent humans from understanding their decision-making mechanisms. This may be contrary to the original intention of human beings to build AI, e.g., giving machines the ability to think like humans. Neuro-symbolic AI which combines the knowledge of symbolic representations with neural networks, seems to

Table 17
A summary of representative concept extraction techniques. KE denotes keyphrase extraction. CE denotes concept extraction. SL denotes statistical learning. Knwl. eng. denotes knowledge engineering. SenticNet denotes the works of Cambria et al. [503,535,536,537]. We do not show the evaluation results for structured concept extraction methods, because they all used very task-specific evaluation methods and datasets, where the results are not comparable.

Task	Reference	Techniques	Feature and KB	Framework	Dataset	Score	Metric
Extractive KE	Zhang et al. [516]	DL	word2vec	Joint-layer RNN	Twitter	86.40%	F1
	Wang et al. [518]	DL	word2vec	BiLSTM, adversarial loss	CCF	29.60%	F1
	Li et al. [519]	Pipeline	word2vec	Similarity matching	MLDBMD	97.00%	MAP
	Al-Zaidy et al. [527]	DL	word2vec	BiLSTM-CRF	KP-20K	35.63%	F1
	Fang et al. [529]	DL	word2vec	Attention; CRF	KP-20K	45.69%	F1
Generative KE	Meng et al. [517]	DL	word2vec	RNN	KP-20K	32.80%	F1@5
	Chen et al. [526]	DL	word2vec	Seq2seq	Krapivin	31.80%	F1@5
	Ye and Wang [532]	DL	word2vec	Seq2seq, semi-supervised	KP-20K	30.80%	F1@5
	Chen et al. [533]	DL	word2vec	Seq2seq, additional Title input	KP-20K	37.20%	F1@5
Structured CE	Havasi and Speer [501]	Knwl. eng.	textual patterns	Pattern matching	–	–	–
	Snow et al. [502]	SL	feature vectors, WN	Probabilistic	–	–	–
	SenticNet	chunking, sem. pars., PoS tag.	syntactic patterns	Syntactic parsing	–	–	–
	Ge et al. [538]	SL	statistics, WN	Elbow algorithm	–	–	–
Clinical CE	Li and Huang [541]	DL	token mention, pos tag, word shape	CNN	TempEval	78.80%	F1
	Liu et al. [542]	DL	word2vec, character2vec	BiLSTM	i2b2–2010	85.78%	F1
	Gehrmann et al. [523]	DL	word2vec	CNN	MIMIC	76.00%	F1
	Yang et al. [543]	DL	word2vec	Transformer	n2c2–2018	88.36%	F1
	Lange et al. [547]	DL	word2vec	Multitask-biLSTM	i2b2–2010	88.90%	F1
Course CE	Pan et al. [524]	Graph	word2vec; Encyclopedia	Graph-based propagation	MOOCs	41.60%	MAP
	Wang et al. [548]	Graph	word2vec, Wikipedia	Graph-based propagation	MOOCs	47.50%	MAP
Patent CE	Liu et al. [505]	ML	self pretrained word2vec, DBpedia	Clustering	USPTO	43.37%	F1

Table 18
A summary of the representative applications of concept extraction in downstream tasks.

Reference	Downstream Task	Feature	Parser	Explain.
Cambria et al. [503]	Sentiment computing	✓	✓	✓
Li et al. [549]	Sentiment computing	✓		✓
Han et al. [550]	Sentiment computing	✓	✓	✓
Xiong et al. [507]	Information retrieval	✓		✓
Liu et al. [551]	Information retrieval	✓		✓
Young et al. [552]	Dialogue systems	✓		
Huang et al. [553]	Dialogue systems	✓	✓	
Fang and Zhang [508]	Commonsense explanation generation	✓		✓

be able to compensate for the lack of model interpretability of pure neural networks because symbolic representations in natural language, e.g., words and concepts are human-readable. We can explain a prediction by viewing what symbolic knowledge is activated. Meantime, symbolic knowledge can represent commonsense knowledge, which is difficult for neural networks to learn from corpora. As the fundamental technique of knowledge base development, concept extraction has a huge potential in downstream applications.

5.9.3. Future works

Open domain concept extraction. Prior research on concept extraction has primarily concentrated on extracting concepts within a particular domain, while other concept extraction efforts aimed at developing knowledge bases have focused on extracting concepts with predefined relations. These approaches severely limit the application scope of knowledge bases. It would be more practical to extract concepts and relations in an open domain, where both the concepts and relations are not focused on specific types. This requires an ontology study to guide the concept extraction, e.g., what can be defined as concepts and relations. It is a more challenging task than the joint NER and relation extraction task, because relationships and concepts are self-aware within a learning model, rather than pre-defined by humans.

Multi-modal concept extraction. "Concept" is also very relevant to human visual recognition. It is argued that for humans, the ability of visual classification is obtained from concept learning, which learned the generalized concept description from sample observations such that a given observation can be identified as a learned concept [555,556]. On the other hand, the abstractness of concepts is strongly related to imagery [557], because abstract concepts are those that are not applicable to tangible, perceptible objects that can be observed through touch, sight, hearing, or other sensory experiences [558]. Thus, learning the relationships between concepts and imagery can help concept extraction research hierarchically organized concepts, e.g., primitives, concepts, and entities. However, till now, there is a lack of research papers working on multi-modal concept extraction to our best knowledge. It could be also interesting to investigate possible synergies in concept extraction between different modalities.

Concept extraction evaluation. Current concept extraction methods were evaluated on an application task, e.g., sentiment analysis to SenticNet or testing specific relationships, e.g., hypernym and hyponym relationship to ConceptNet and WordNet extension. The issue with such an evaluation method is that it can only reflect the effectiveness of a developed knowledge base or concept extraction method on a specific domain. Since different knowledge bases have different application targets, it's hard to evaluate and compare them with unified criteria. It would be valuable to propose a framework for knowledge base evaluation that is independent of specific tasks. It would be helpful to understand the quality of included concepts, relationships, and their representations.

More concept extraction applications. Despite the attention some scholars have given to neuro-symbolic AI, the body of related works remains relatively scant in comparison to end-to-end neural network models. One possible explanation for this disparity is that, at present, there is greater emphasis placed on the accuracy of the model rather than the transparency of its decision-making process. Thus, there is a need for more concept extraction applications, which can aid in enhancing the explainability of neural network-based models. It offers insights for the development of knowledge bases, prompting researchers to reassess how they extract and organize concepts in order to more effectively support subsequent applications.

6. Subjectivity detection

Conventionally, subjectivity detection is defined as a task to determine whether a text is subjective or not, where a subjective text expresses personal feelings, evaluations, and speculations [559], whereas an objective one merely delivers factual information. Generally, subjectivity can manifest in different forms, e.g., opinions, allegations, desires, beliefs, and suspicions [560] to express private states. It is not an easy task to identify the use of subjective language, as a subjective sentence does not always contain an opinion [560]. Therefore, it is important for the subjectivity detection task to find reliable clues. Aside from opinion-bearing words, syntax also provides essential clues in reporting private states, because grammaticalization involves the recruitment of items to mark the speaker's point of view [561].

Early works often equated the presence of subjectivity to the presence of subjectivity-bearing words in a sentence [562–565]. However, subjectivity is context- and domain-dependent. Some words are only subjective in certain contexts or domains. Therefore, many researchers incorporated syntactic dependencies [566,567], interactions between neighboring sentences [559,568] or in discourse [569] to extract different levels of contextual information. An alternative to this subjective-lexicon-based approach is the word-frequency-based approach [570,571], which is completely domain-independent by learning from document-level information. However, this approach has difficulties capturing syntactic dependencies. By now, subjectivity detection research has been divided into several distinct tasks, each with its unique objectives. One such task is individual subjectivity detection, which focuses on detecting subjectivity at the sentence level. In contrast, context-dependent subjectivity detection aims to incorporate discourse information and a broader context in detecting subjectivity. Cross-lingual subjectivity detection, on the other hand, strives to identify subjectivity in various languages. Moreover, multi-modal subjectivity detection is concerned with identifying subjective expressions in different modalities such as audio and video. Finally, the bias detection task is centered on identifying biased statements in ostensibly impartial articles.

Subjectivity detection is commonly considered as a sub-task of sentiment analysis since it serves as a filtering step for polarity detection [560]. It can also be helpful for downstream tasks that require a distinction between opinionated and non-opinionated sentences, such as opinion and information retrieval [572,573], analyses in financial and political domains [574–576], question answering systems [577,578], etc.

6.1. Theoretical research

Given the broadness of subjective expressions, e.g., expressing personal feelings, evaluations, and speculations, the related theoretical research in this domain is also rich.

6.1.1. Subjective elements

Early linguistic works studied subjective language extensively in third-person narrative text. Banfield [579] defined the *SELF* of a sentence as the speaker in conversation or the narrating character in third-person fictional text. She identified a variety of morphological, lexical, and syntactic elements, termed subjective elements, that always express the private states, i.e., emotions and opinions, of the sentence's *SELF*. However, many linguistic elements are subjective only in certain conditions. Therefore, following [579], [580] further defined a category termed potential subjective elements, which expanded the subjective elements with some linguistic elements that can, but not always, report the private state of a character. Wiebe et al. [581] applied these findings to identify the subjective language in the non-fictional text, suggesting that potential subjective elements are also valid subjectivity clues for texts other than third-person narrative fiction.

6.1.2. Speech acts

Speech acts have a strong connection with subjective expressions because speech acts perform actions, such as making a promise, giving an order, or expressing a belief. Austin [582] argued that language is not just a tool for describing the world but also a means of accomplishing things in the world. Through speech acts, individuals can influence the world around them and the actions of others. In this sense, many seemingly objective expressions with speech acts can become subjective. For example, if someone says,

(18) I promise to do it.

The utterance is not just conveying information but also performing the act of making a promise. A more subjective case is

(19) I believe that it will rain tomorrow.

When individuals express belief in such a manner, they are essentially asserting their mental disposition or perspective towards a specific statement. This entails making a claim about their inner state or outlook toward a proposition. Austin [582] argues that a considerable number of utterances possess illocutionary force, which signifies that their purpose is not merely to communicate information but also to accomplish something beyond that. Thus, subjective expressions may be more than we think in our everyday language.

6.1.3. Conceptual metaphor

Lakoff and Johnson [539] argued that metaphors are not solely a linguistic phenomenon, but also mirror human cognition via concept mappings. When an individual uses a metaphorical expression, they employ a source concept to represent a target concept in a particular context, thereby conveying their cognitive attitude toward the target concept. This process, known as concept mappings, facilitates such representation. In instances such as the statement

(20) Our love is a journey.

The individual utilizes the concept of a "journey" as the source to represent the target concept of "love", expressing their subjective feeling that their love is characterized by both ups (joy) and downs (sadness). "Our love is a journey" cannot be an objective statement, because the two concepts are from different domains, i.e., literally, love is not a journal. Thus, there is a semantic contrast between the literal and contextual meanings of a metaphor [583]. The semantic disparities inherent in metaphors suggest that relying on the literal meanings of a statement alone is insufficient in substantiating its subjectivity. Even though the statement of List (20) does not use any obvious opinionated words, e.g., “happy” and “sad”, it also expresses a personal feeling. Thus, the pragmatics of statements must also be taken into account in subjective detection.

6.2. Annotation schemes

For general subjectivity detection, it is sufficient for a dataset to annotate a sentence, snippet, or document as subjective (positive/negative) or objective (neutral). Nevertheless, [584] proposed the MPQA scheme, which annotates text at the word and phrase levels. The MPQA scheme is suitable for fine-grained subjectivity detection that aims to identify the source, target, and properties of each expression of the private state.

Wilson [585] proposed the AMIDA Scheme for annotating subjectivity in speech. This scheme marks word spans that are in the following three main categories: subjective utterances, objective polar utterances, and subjective questions. A subjective utterance is a word span that expresses a private state. An objective polar utterance delivers positive or negative factual information without expressing a private state. A subjective question is a question in which the speaker is eliciting the private state of someone else. Each category is divided into finer classes that indicate the polarity and certainty of an utterance.

6.3. Datasets

A summary of all the introduced datasets can be found in Table 19. Generally, subjectivity detection data are organized in the following forms. A text is typically labeled as either subjective or objective, with the former category often further classified as positive, negative, or neutral. The following examples are from SemEval-2013 Task 2B: Sentiment Analysis on Twitter [586].

```
id1: "264215390773727232"
id2: "276151090"
text: "Alex Poythress had 11 points and 7 rebounds in his debut
with Kentucky during an exhibition game on Thursday. He played
28 minutes."
label: objective

id1: "263732569508552704"
id2: "369152026"
text: "Kick-off your weekend with service! EV!'s Get on the Bus
trip to the Boys &amp; Girls Club is Friday from 3-6! Hope to see
you there :)"
label: "positive"

id1: "213342054351257601"
id2: "189656827"
text: "Desperation Day (February 13th) the most well known day in
all mens life."
label: negative

id1: "263803288074477568"
id2: "396953010"
text: "It seem like Austin Rivers is tryin to had to get a bucket.
I feel em tho my 1st game in the league I was trying hard too"
label: neutral
```

For fine-grained subjective annotation, the labels are annotated at the span level. The following examples are from SemEval-2013 Task 2A: Sentiment Analysis on Twitter [586].

```
id1: "255732290246815744"
id2: "315400337"
text: "Billy Cundiff may be leaving Washington. Hopefully he
won't miss the door on the way out."
start id: "7"
end id: "7"
label: "positive"

id1: "255732290246815744"
id2: "315400337"
text: "Billy Cundiff may be leaving Washington. Hopefully he
won't miss the door on the way out."
start id: "9"
end id: "10"
label: "positive"
```

MultiParty Question Answering (MPQA) [584] is derived from 535 English news articles from a wide variety of news sources, manually annotated for subjectivity. The corpus contains 9700 sentences, 55% of which are labeled as subjective and 45% as objective. The MPQA Gold [587] contains 504 Spanish sentences manually annotated for subjectivity, where 273 sentences are subjective and 231 are objective. The Multi-MPQA [588] contains parallel corpora to the MPQA dataset in five languages other than English, namely, Arabic, French, German, Romanian, and Spanish.

The Movie Review dataset (Movie) [568] contains 5000 movie review snippets collected from Rotten Tomatoes,[34] considered as subjective. Furthermore, 5000 sentences are collected from plot summaries from the Internet Movie Database (IMDB),[35] considered as objective. All reviews and plot summaries are sourced from movies released post-2001, preventing overlap with the polarity benchmark dataset [568]. A data sample, either sentence or snippet, is at least 10 words long. The Debate dataset [576] is derived from the political and ideological dataset [589], containing 53,453 sentences from political and ideological posts and comments. The instances are automatically labeled for subjectivity by using lexicon-based and syntactic-pattern-based classifiers [562].

Numerous microtext corpora exist that can serve as benchmark datasets for subjectivity detection. Barbosa and Feng [590] presented a dataset containing 200,000 English tweets, where roughly 100,000 are subjective and the rest are objective. Serrano-Guerrero et al. [591] manually annotated 498 English tweets as positive, negative, or neural. SemEval 2013 [586] is a collection of 12,002 English tweets labeled as objective, positive, negative, or neutral. Nuclear Energy Tweets (NET) [592] contains 2308 English tweets about nuclear energy, manually annotated for subjectivity. The Multilingual Tweets (MLT) dataset [593] is a collection of 12,719 tweets about nuclear energy in English, French, Spanish German, Malay, and Indonesian, 7700 out of which are manually labeled for subjectivity. The Taller de Analisis de Sentimientos en la SEPLN (TASS) corpus [594] contains 10,000 tweets in Spanish, collected from posts by 150 public figures in fields of sports, politics, and communication during the period from 2011 to 2012. Each tweet is labeled as positive, neutral, negative, or without opinion.

The Web Document dataset [595] contains 1076 English web documents, sourced from traditional news websites and blog posts on diverse topics. Each document is manually annotated as objective, positive, or negative. The Text REtrieval Conference (TREC) dataset [577] is a collection of 8000 WSJ articles evenly distributed in the categories of editorial, letter to editor, business, and news. The articles and sentences from the former two categories are mapped as opinions (subjective), while the ones from the latter two are facts (objective). The Forum

[34] https://www.rottentomatoes.com

[35] https://www.imdb.com

Table 19
Subjectivity detection datasets and statistics. ISD denotes individual subjectivity detection. CDSD denotes context-dependent subjectivity detection. CLSD denotes cross-lingual subjectivity detection. MMSD denotes multi-modal subjectivity detection. BD denotes bias detection.

Dataset	Task	Source	# Samples	Reference
MPQA	ISD, CDSD	English news articles	9,700	Wiebe et al. [584]
MPQA Gold	ISD, CLSD	Spanish sentences	504	Mihalcea et al. [587]
Multi-MPQA	ISD, CLSD	Machine-translated MPQA	9,700	Banea et al. [588]
Movie	ISD, CDSD	Rotten tomatoes, IMDB	10,000	Pang and Lee [568]
WebDoc	CDSD	English web documents	1,076	Chesley et al. [595]
TREC	CDSD	WSJ	2,000	Yu and Hatzivassiloglou [577]
Debate	CDSD	Political and ideologicaldataset	53,453	Al Hamoud et al. [576]
Twitter1	ISD	English tweets	200,000	Barbosa and Feng [590]
Twitter2	ISD	English tweets	498	Serrano-Guerrero et al. [591]
Forum	CDSD	Online forums	700	Biyani et al. [569]
SemEval 2013	ISD	English tweets	12,002	Nakov et al. [586]
NET	ISD	English nuclear energytweets	2,308	Khatua et al. [592]
MLT	ISD, CLSD	Multi-lingual nuclearenergy tweets	7,700	Satapathy et al. [593]
TASS	ISD, CLSD	Spanish tweets	10,000	Villena et al. [594]
Email	CDSD	BC3 corpus	1,800	Murray and Carenini [596]
AMIDA	MMSD	AMI Meeting Corpus	13	Wilson [585]
ICT-MMMO	MMSD	Youtube review videos	370	Wöllmer et al. [599]
MOUD	MMSD	Youtube review videos	498	Morency et al. [600]
Conservapedia	BD	Conservapedia statements	1,000	Hube and Fetahu [601]
WNC	BD	Wikipedia sentence pairs	180,000	Pryzant et al. [602]

Table 20
Useful knowledge bases for subjectivity detection.

Name	Knowledge	# Entities	Structure
The General Inquirer	Sentiment labels	4,000	List
MPQA Subjectivity Lexicon	Subjectivity clues	8,000	List
SentiWordNet	Structured lexical knowledge by concept	100,000	Graph
WordNet-Affect	Lexical knowledge	4,787	Graph
SenticNet	Sentiment scores	200,000	Graph

dataset [569] contains 700 threads from online forums Trip Advisor–New York[36] and Ubuntu Forums,[37] manually annotated for subjectivity. Email [596] contains 1800 sentences derived from BC3 corpus [597], 172 out of which are labeled as subjective.

For multi-modal subjectivity detection, the AMIDA dataset [585] consists of 19,071 dialogue act segments from 20 conversations from the AMI Meeting Corpus [598], manually annotated with the AMIDA scheme. 42% of the dialogue act segments are tagged with at least one subjective annotation. The Institute for Creative Technologies Multi-Modal Movie Opinion (ICT-MMMO) dataset [599] contains 370 Youtube review videos labeled as strongly negative, weakly negative, neutral, weakly positive, and strongly positive. Multimodal Opinion Utterances Dataset (MOUD) [600] is a collection of 80 Youtube review videos annotated as positive, negative, and neutral.

For the bias detection task, which aims to identify subjective bias in Wikipedia, the following datasets are widely used. Conservapedia [601] is a collection of 1000 single-sentence statements from Conservapedia,[38]manually annotated as biased or unbiased. Wiki Neutrality Corpus (WNC) [602] contains 180,000 aligned Wikipedia sentence pairs. Each pair consists of a sentence before and after bias neutralization by English Wikipedia editors.

6.4. Knowledge bases

Lexicons of subjectivity clues and patterns are commonly used for subjectivity detection, as summarized in Table 20. The General Inquirer [603] is a lexicon consisting of 10,000 words sorted into 180 categories for content analysis. The Subjectivity Clues lexicon [562] is a list of words that are subjective in most cases (strongly subjective) and words that may have subjective use in certain contexts (weakly subjective). MPQA Subjectivity Lexicon [604] expanded the Subjectivity Clues using additional dictionaries and lexicons, containing over 8000 subjectivity clues.

Knowledge bases that provide sentiment information are also widely used for subjectivity detection. WordNet-Affect [605] is a set of synsets derived from WordNet that effectively represents affective concepts. SentiWordNet, as introduced in the previous section, is based on WordNet. Each word in SentiWordNet is given three scores indicating its positivity, negativity, and objectivity. SenticNet [503] is a concept-level knowledge base that includes semantic, sentic, and polarity associations.

6.5. Evaluation metrics

The performance of subjectivity detection is commonly evaluated via accuracy and F-measure.

6.6. Annotation tools

The aforementioned NER annotation tools (see Section 4.6) can be used for subjectivity detection because these tools can annotate labels for spans (fine-grained subjectivity detection) and sentences (coarse-grained subjectivity detection).

6.7. Methods

6.7.1. Individual subjectivity detection

In individual subjectivity detection, the subjectivity of a sentence is evaluated in isolation and irrespective of any contextual factors. The primary methods used for addressing this task include lexicon-based, word frequency, and deep learning approaches.

A. Lexicon-based

Drawing on the premise that sentences that contain commonly-subjective expressions are more likely to be subjective, lexicon-based methods utilize a manually-constructed lexicon of subjective words, clues, or patterns to determine the subjectivity of a given sentence.

Riloff and Wiebe [562] introduced an unsupervised rule-based classifier that leverages the identification of subjective clues and patterns to detect subjective sentences, while also employing bootstrapping to recognize objective sentences based on the absence of such indicators. The clues were manually collected and annotated. The patterns were generated by the AutoSlog-TS algorithm [606], based on pre-defined

[36] http://www.tripadvisor.com/ShowForum-g60763-i5-New_York_City_New_York.html
[37] http://ubuntuforums.org
[38] http://www.conservapedia.com

syntactic templates. Wiebe and Riloff [607] further improved this bootstrapping system by using the labeled sentence produced by the rule-based method as initial training data for a Naïve Bayes classifier. The major weakness of these methods is the unreliable assumption that the absence of subjective clues and patterns indicates objectivity, resulting in false-positive errors.

Kim and Hovy [563] first compiled lists of words that convey opinions and those that do not, which were manually annotated with corresponding classes and levels of strength. They expanded the lists with a common English word list by measuring the WordNet distance between a common word and the compiled seed lists. They further identified additional opinion words and non-opinion words from editorial and non-editorial WSJ documents by computing their relative frequencies. By detecting the subjectivity of a given sentence based on the presence of a single strong valence word, their method achieved 65% accuracy on MPQA.

Benamara et al. [608] argued that sentence-level subjectivity detection cannot fully leverage context, because a sentence may contain several opinion clauses, and opinion expressions may be discursively related. As such, they proposed a segment-level annotation based on the Segmented Discourse Representation Theory [609], where segments are labeled as explicitly subjective, implicitly subjective, subjective non-evaluative, and objective. This fine-grained annotation can better enhance polarity detection, as segments in the latter two categories do not covey positive, negative, or opinion. However, the limitation of this method is that the four label classes are unbalanced in the corpus. Additionally, implicitly subjective segments are often nuanced and hard to identify. Thus, it would be challenging to design an appropriate classifier. The paper circumvented this problem by reframing the task as two parallel binary classification tasks and obtained 82.31% accuracy with a manually compiled French lexicon and SVMs as classifiers.

Merely detecting the existence of subjective keywords is often an insufficient indication of a sentence's subjectivity. Other works attempted to enrich the feature set by incorporating more sentence-level information. Relying on expert knowledge of parse tree, [567] manually constructed a set of syntax-based patterns from unigrams and bigrams to extract features. A MaxEnt model was employed as the classifier, obtaining 92.1% accuracy on the Movie dataset. Remus [610] hypothesized that the readability of a sentence was related to its subjectivity. Hence, readability formulae such as Devereux Readability Index [611] and Easy Listening [612] were incorporated as features in addition to the MPQA Subjectivity Lexicon, obtaining 84.5% F-measure on Moive.

Compared to standard text, microtext such as tweets contains informal and irregular expressions, making it more difficult for machines to process. Many works proposed subjectivity detection systems that specifically targeted Twitter text. Given the word constraint imposed by Twitter, a tweet is generally regarded as a sentence. Barbosa and Feng [590] believed that using subjectivity detection as an upstream task would improve the performance of polarity detection on Twitter text. Aside from conventional features such as subjective clues and PoS tags, they leveraged Tweet-specific syntax features, e.g., links and upper case. An SVM classifier was employed, which achieved 81.9% accuracy on the Twitter dataset, and improved the accuracy of polarity detection by 5.6%. Following their footsteps, [613] incorporated more Tweet-specific features that leveraged the structure of Twitter, e.g., the relationship between tweets, users, hashtags, and links. Using the stacking classifier proposed by Cotelo et al. [614], their method obtained 89.8% on TASS. To reduce human effort, [615] created a Twitter subjectivity lexicon automatically through a meta-heuristic approach, i.e., a genetic algorithm, which produced separate lists of subjective and objective words. A Bayse network was employed to classify a given tweet based on its subjective and objective word counts, achieving 60.9% on SemEval 2013. Alternatively, [592] leveraged the concept-level knowledge base SenticNet as their lexicon, which is able to provide implicit meaning associated with commonsense concepts. Their method obtained 80.7% accuracy on the NET dataset.

The methods introduced above have a common limitation, i.e., the lexicons are lists of keywords, instead of word meanings. Some subjective clues in fact have both subjective and objective word senses, which are not distinguishable in keyword lexicons, leading to false-positive errors. This problem can be mitigated by incorporating a Subjectivity WSD (SWSD) system to build a sense-aware lexicon. Akkaya et al. [616] trained a supervised targeted SWSD system using SVM. The training data was compiled using words that are both in the MPQA Subjectivity Lexicon and the sense-tagged SENSEVAL corpora [617–619]. Alternatively, [620] applied an unsupervised, clustering-based SWSD system [621] on SentiWordNet to label each subjective word with fine-grained sense. Both SWSD systems were applied to a rule-based classifier similar to the one proposed by Riloff and Wiebe [562]. The supervised one improved accuracy by 1.3% on MPQA, while the unsupervised one improved F-measure by 6.5% on Movie. A prominent limitation of lexicon-based methods is that they require external resources such as sentiment lexicon and knowledge base.

B. Word Frequency

Word-frequency-based methods detect subjectivity by modeling word presence or occurrence within a corpus. Therefore, compared to lexicon-based methods, they are language-independent and require neither manual annotation nor linguistic knowledge. They are also less computationally expensive due to the reduction of feature sets.

Rustamov et al. [570]; Kamil et al. [571] proposed a language-independent feature extraction algorithm with a novel statistical measure of word occurrence called Pruned ICF (Inverse-Class Frequency), which is proven to be more effective than the standard IDF (Inverse-Document Frequency). Additionally, they applied two widely-used methods for pattern recognition to detect subjectivity, namely Fuzzy Control System (FCS) [622] and Adaptive Nero-Fuzzy Inference System (ANFIS) [623], achieving the accuracy of 91.3% and 91.66% on the Movie dataset, respectively. The latter obtained better performance due to the addition of a neural network layer. Inspired by empirical evidence that hybrid systems improve the performance of NLP classifiers, [624] further integrated FCS, ANFIS, and HMM into a sequential hybrid system, where input sentences that are wrongly labeled by the prior classifier are passed onto the subsequent one. Using the same feature extraction method as the previous paper, this system increased the accuracy to 92.24% on Movie.

Wang and Manning [625] proposed a novel dropout algorithm to optimize the feature learning process. Conventional dropout training in neural network [626] prevents feature co-adaptation by randomly sampling neurons and input features and setting them to zeros, which leads to slow training. The authors suggested fast dropout training as a more efficient alternative, using a Gaussian approximation to draw samples. They applied this dropout method to Naïve Bayes Support Vector Machine (NBSVM) [627], which extracts features based on word presence. Their method not only achieved the accuracy of 93.6% on Movie and 86.3% on MPQA, but also greatly decreased the training time. Experiments also showed that fast dropout training could be applied to other loss functions and neural networks.

Latent Dirichlet Allocation (LDA) Blei et al. [67] is a weakly-supervised generative model that assumes every document is a distribution of latent topics, which is determined by word frequencies. He [628],Maas et al. [629] suggested that subjectivity detection can be solved by LDA, based on the intuition that subjective sentences likely contain opinionated words. Hence, the paper modified conventional LDA so that the latent topics are word-level sentiment labels. An additional layer is inserted between word and document levels to model sentence-level subjectivity labels. Sentiment lexicons are incorporated to establish an informed prior distribution for word-level sentiment labels, achieving 71.2% accuracy on MPQA. On the other hand, [630] argued that LDA likely discovers topics based on semantic similarities, instead of sentiment. Therefore, they modified LDA, so that it directly models word probabilities conditioned on topic distributions to capture semantic information. To explicitly extract sentiment information, they incorporated supervised sentiment analysis as an auxiliary task. Their method achieved 88.58% on the Movie dataset. The drawback of the word frequency approach is that the order of the words is not considered. Thus, syntactic information cannot be effectively learned using this approach.

C. Deep Learning

The acquisition of precise sentence representations is crucial for subjectivity detection, and as such, numerous studies have examined neural sentence modeling as a language-independent alternative to parse trees. Kalchbrenner et al. [631] presented a Dynamic CNN (DCNN) that is able to capture short- and long-range relations. The core component of DCNN is dynamic pooling, which outputs the sub-sequence of k maximum values in the input sequence, where k can be dynamically chosen. Hence, DCNN produces a hierarchical feature graph that contains syntactic, semantic, and structural patterns of the input sentence. However, their sentence representations do not retain any intermediate information, e.g., word-level and phrase-level features. To address this, [632] described a self-adaptive hierarchical sentence model named AdaSent. Inspired by gated recursive CNN [633], AdaSent forms a pyramid-shape directed acyclic graph, where the bottom level is word representations and the top level is sentence representations. In this process, the gating network receives information from each level and selects the most appropriate representations for the given task. Their method obtained an accuracy of 95.5% on Movie and 93.3% MPQA.

With similar motivation for higher-order dependencies, [634] proposed a Bayesian Network-based Extreme Learning Machine (BNELM) framework for subjectivity detection. Single-layer feedforward neural networks, known as Extreme Learning Machines (ELMs), excel at inductive learning. However, the excessive number of hidden neurons in ELMs often leads to overfitting and slow performance. To address these weaknesses, Bayesian networks were introduced to model connections among the hidden neurons of ELM, as they can prune redundant and irrelevant hidden neurons and capture high-dimensional features. Furthermore, ELM cannot handle non-linear data such as sequences of sentences. Thus, an RNN layer was used to extract temporal features. Upon it, a fuzzy classifier was applied to achieve stability in case of noisy data, producing the output labels. Additionally, a deep CNN was employed prior to BNELM to provide low-dimensional features. The framework achieved the accuracy of 75% on MPQA Gold and 89% on TASS, outperforming previous ELM-centric models, namely, standard ELM and Sparse Bayesian ELM [635].

Likewise, [593] employed CNN and RNN to extract spatial and temporal information respectively. To make the model more robust, they incorporated reinforcement learning, namely Point-wise Probability Reinforcement (PPR) [636], to regularize the learning process of CNN and reduce the influence of outliers. Specifically, convolutional layers in the CNN component were added iteratively, where the weight of each neuron was fine-tuned by the reinforced maximum likelihood of PPR. Their method did not perform very well on MPQA, obtaining 50% F-measure. However, it achieved a good performance of 76% F-measure on the multi-lingual Twitter dataset MLT.

In the same vein, PLMs can also provide beneficial universal representations learned from a plethora of unlabeled text. For instance, Al Hamoud et al. [576] fed GloVe embeddings to different types of RNN variants, among which LSTM with attention mechanism achieved the best accuracy of 89.53% on MPQA and 83.83% on their proposed political and ideological dataset, whereas Bi-LSTM with attention achieved the best accuracy of 92.8% on Movie. Kim [637] fine-tuned pre-trained Word2Vec with a simple CNN, obtaining accuracy of 93.4% on Movie and 89.5% on MPQA.

Furthermore, many works observed that it is complementary to combine PLM and MTL for more effective learning of text representations [638–640]. Motivated by this, [641] fine-tuned BERT using MTL, where the BERT layers are shared among subjectivity detection and three other text classification tasks. Similarly, [642] proposed an MTL framework for subjectivity and polarity detection. The framework leverages BERT as embedding, which is fed into two separate self-attention Bi-LSTM layers. A neural Tensor Network (NTN) [643] was used as the information-sharing layer. Both methods employed a simple softmax classifier for each task. The former achieved 95.23% accuracy on Movie, while the latter obtained 95.1%. However, a shared limitation is that, despite their overall good performance, some of the tasks did not exceed single-task learning baselines. This is likely because both methods adopted hard parameter sharing MTL [644], which emphasizes more on generalization rather than optimization.

Sagnika et al. [645] presented an attention-based CNN-LSTM model for subjectivity detection, which served as a pre-processing step for sentiment analysis. The combination of CNN and LSTM enabled the model to capture both spatial and temporal information. Additionally, it utilized word embeddings enhanced by sentiment-related information [646]. Initially, the training of the model was carried out with the Movie dataset, after which it was utilized to analyze the sentiment of the IMDb dataset. The objective sentences were eliminated from the dataset to form a modified set of reviews. Various models were tested as sentiment classifiers. The subjectivity detection model not only obtained 97.1% accuracy on the Movie dataset, but also consistently improved the performance of sentiment analysis.

6.7.2. Context-dependent subjectivity detection

The method of individual detection categorizes each sentence without considering its context. However, subjectivity detection and sentiment classification are contextual problems since lexical items can affect each other in a discourse setting [647,648]. Pang and Lee [568] was the first to leverage inter-sentence context information to filter out objective sentences, in order to better serve document-level polarity detection. Based on the hypothesis that adjacent text spans might have the same subjectivity label [559], suggested an algorithm known as the “minimum cuts algorithm” that aims to optimize the subjectivity status score for every sentence separately, while also punishing the assignment of different labels to two closely related sentences. These two sub-objectives are independent of each other, making the model more flexible for the addition of features.

Context-dependent methods can be divided into two categories, namely, the feature engineering approach and the statistical approach.

A. Feature Engineering

A common way to incorporate document-level information is by designing relevant features. Das and Bandyopadhyay [649] proposed a domain-independent rule-based algorithm, named theme detection. The model utilized document-level features, e.g., positional aspects (document title, first paragraph, last two sentences), the positions of subjectivity clues, and the distance between any two thematic words. As with many techniques at the sentence level, this approach also integrated syntactic characteristics and resources such as SentiWordNet and MPQA Subjectivity Lexicon. It achieved precision and recall of 76.08% and 83.33% on MPQA.

To automatically select an appropriate feature set, [650] employed the genetic algorithm (GA) [651,652], which is a probabilistic search method, to find the optimal range of values of every feature. To capture context information, positional aspects, word distribution, and document theme [581] were incorporated as discourse-level features, aside from the commonly-used lexical and syntactic features. The GA then identified the globally optimal feature set by natural selection and computed the corresponding accuracy of the classifier through the fitness function. An advantage of the proposed method over other statistical classifiers is that the entire input sentence is encoded by GA and used as features, instead of using n-gram. Their method obtained the F-measure of 93.02% on MPQA and 95.69% on Movie.

Biyani et al. [569] noticed a gap in subjectivity detection targeting online forums. Moreover, they argued that lexical features are highly dimensional, leading to the risks of overfitting and slow training. Thus, they presented a Forum dataset, and designed a set of non-lexical thread-specific features. Specifically, they leveraged thread structure and dialog acts and utilized lexicons and tools such as MPQA Subjectivity Lexicon and SentiStrength [653] to extract sentiment features. With the addition of conventional lexical features, the logistic regression classifier obtained 77.01% accuracy on Forum.

B. Statistical Approach

To minimize human effort in designing features, a statistical approach automatically learns features from a given corpus using statistical models. Yu and Hatzivassiloglou [577] simply implemented a Naïve Bayes classifier for document-level subjectivity detection, which achieved the F-measure of 97% on the TREC dataset proposed by them. Motivated by the observation that language models are adept at representing knowledge of the text they were trained on, [654] proposed a language-model-based document-level subjectivity detection method. During training, a subjective reference language model and an objective were built using labeled documents. During inference, a language model was constructed for each input document, which was compared with the reference language models using KL-divergence [655], producing two similarity scores. The difference between these two scores was regarded as the subjectivity score of the document. The final output of the model was a sorted list of input documents, based on their subjectivity scores. To achieve language non-specificity, the paper also proposed a semi-supervised method where the reference language models were built on a lexicon divided into subjective and objective parts, based on polarity scores. The supervised method obtained 94.63% MAP on the Movie dataset, whereas the unsupervised obtained 53.61% MAP.

Word embeddings can only provide limited syntactic and semantic information [656]. Therefore, to better initialize their model, [657] employed a Gaussian Bayesian Network (GBN) [658] layer to capture long-range features among successive sentences, which were used to pre-train the CNN classifier. The GBN layer converted the sentence sequence from the MPQA dataset into a time series of word frequency, captured second-order word dependencies with a time lag of 2, and generated a subset of sentences that contained the most significant words from the MPQA Subjectivity Lexicon. The model adopted a CNN sentence model with convolution kernels of increasing size, which combined the local word dependencies within the kernel size to model long-range syntactic relations. It was pre-trained with the sub-set of sentences produced by GBN before being trained on the full dataset, obtaining the accuracy of 93.2% on MPQA and 96.4% on Movie.

6.7.3. Cross-lingual subjectivity detection

A. Language-Independent Approach

For feature-engineering-based subjectivity detection, lexical resources and tools are often not readily available for non-English languages. A common approach to circumvent this problem is to use non-language-specific features that are based on the presence or occurrence statistics of a corpus, e.g., word frequency [67,570,571,625,630,656] and language modeling [654]. Mogadala and Varma [659] further introduced language-independent feature weighing, leveraging unigram and bigram frequencies, and unigram word length. Entropy-based category coverage difference [660] was employed as the feature selection method.

B. Translation Approach

Another solution is the translation approach, where lexical resources for the target language are automatically generated by translating the resources and tools available for English, usually with the help of statistical machine translation (SMT) [587,661–665]. Banea et al. [588] conducted a study on English and five other highly lexicalized languages, proving that a multi-lingual feature space constructed through SMT improved the accuracy of subjectivity detection on all languages involved. However, the sentence translation process can lead to the loss of essential lexical information such as inflection and formality, which often served as an indicator of subjectivity [662]. Chaturvedi et al. [666] mitigated this information loss during translation by using a neural network to transfer resources from English to Spanish. They first translated the MPQA Subjectivity Lexicon into Spanish using an SMT system [667]. A MaxEnt-based PoS tagger [668] and a multi-lingual WSD system [669] were incorporated in the preprocessing stage to minimize the loss of lexical information during translation. Their proposed model, named Lyapunov Deep Neural Network (LDNN), extracted spatial features from the input Spanish sentence and its translated English form using CNN, which were then combined with an RNN to capture the bilingual temporal features. To mitigate the vanishing gradient problem with RNN, a Lyapunov function was used as the error function of RNN for stable convergence. Utilizing the high-level features produced by Lyapunov-guided RNN, a multiple kernel learning [670,671] classifier yielded the prediction. Their model obtained 84.0% F-measure on MPQA Gold, and 88.4% accuracy on TASS.

6.7.4. Multi-modal subjectivity detection

While most studies on detecting subjectivity have concentrated on text-based data, the identification of subjective expressions in other modalities, such as audio and video, presents an important area for research. For instance, [596,672] proposed an automatic pattern extraction method for subjective expression in spoken conversation, which is able to extract Varying Instantiation N-Grams (VIN) from labeled and unlabeled data. Unlike convention n-gram, a VIN is a trigram where each unit can be either a word or a PoS label, which is a more robust alternative to syntactic parsers for fragmented and disfluent text, such as meeting transcripts. Combined with a large raw feature set, a MaxEnt classifier scored the F-measure of 52% on the AMIDA dataset.

The method above, however, did not leverage any information from other modalities. Raaijmakers et al. [673] explored the effectiveness of lexical and acoustic features in speech subjectivity detection. Specifically, they investigated word, character, prosody, and phoneme n-grams. Following [674,675], the prosodic features were extracted based on pitch, energy, and the distribution of energy in the long-term averaged spectrum. The word-, character-, and phoneme-level features were extracted from manual speech transcripts. A separate BoosTexter classifier [676] was employed for each feature set, whose predictions were combined using a simple linear interpolation strategy [677] to obtain the final output. The combination of the four types of feature sets achieved 75.4% accuracy and 67.1% F-measure on AMIDA. Furthermore, experiments showed that word- and character-level features contributed the most to higher results, whereas prosodic features yielded marginal improvements.

6.7.5. Bias detection

Bias detection refers to the task of identifying biased statements from supposedly impartial articles. Specifically, in Wikipedia, the Neutral Point of View (NPOV) is a core principle that ensures neutrality for controversial topics. Thus, the goal of this task is to detect sentences that violate NPOV policy on a Wikipedia page. Bias detection is closely related to subjectivity detection. Its development mirrors the technical trends of the latter. However, it is considered to be more complex, because the linguistic cues of biased language are often nuanced, and depend heavily on the context.

For lexicon-based approaches, [678] manually compiled a biased word lexicon and feature set that covered framing bias (use of subjective words or phrases that links to a particular point of view), and epistemological bias (linguistic cues that modify the credibility of a statement). However, their method focused only on detecting a single bias-inducing word in a known biased statement. Furthering their work, [601] constructed a more comprehensive biased word lexicon for sentence-level bias detection. To minimize human efforts, they leveraged Word2Vec to expand a seed word list by measuring the distance between word vectors. Aside from the lexicon, other syntactic and semantic features were incorporated, e.g., tri-gram, PoS tags, Linguistic Inquiry Word Count (LIWC) [679], framing bias features, and epistemological bias features. By using a Random Forest classifier, their method obtained 74% precision on their proposed Conservapedia dataset. Aleksandrova et al. [680] proposed a semi-automatic method to construct a multi-lingual bias detection corpus, consisting of Bulgarian, French, and English sentences from Wikipedia. Their method was applicable for building a corpus from a Wikipedia archive in any language, as it does not rely on language-specific features. Additionally, they provided the performance of three baseline models, namely BoW, fastText [681], and logistic regression [682], among which BoW achieved the best overall average F-measure of 59.57% across the three languages.

For neural network approaches, [683] employed RNN to capture the inter-dependency of words and their context. To address the weakness of RNN in modeling long-range information, a hierarchical attention mechanism [684] was adopted, which applied word-level attention on each sentence to compute sentence representations, upon which sentence-level attention was applied to learn biased cues from different samples. Following previous feature-based works, they concatenated GloVe embedding, PoS tags, and LIWC features as word representations.

PLMs were also widely used in bias detection. Pryzant et al. [602] extended the work of Recasens et al. [678] by using a pre-trained BERT-based detector to identify bias-inducing words and neutralizing them via an LSTM-based editor. A join embedding mechanism was employed to allow the detector control over the editor. They also introduced the WNC dataset for detecting and editing biased language, on which their model obtained 93.52% BLEU and 45.80% accuracy for the produced edits. However, a limitation is that they primarily targeted single-biased words. To mitigate this, [685] enabled multi-word detection by identifying bias at the sentence level. They employed the weighted-average ensemble method on several BERT-based models to detect biased language, which obtained 71.61% accuracy and 70.40% F-measure on WNC.

6.8. Downstream applications

6.8.1. Sentiment computing

The presence of objective texts can dilute the task of sentiment computing. Therefore, the machine can better classify the remaining non-objective opinions by using subjectivity detection as an upstream task [593,686]. For document-level sentiment analysis specifically, [687] showed that subjectivity detection reduced the amount of data to 60% while still producing the same polarity classification results as full-text classification. The analysis reveals that a considerable portion of real-world textual data is objective in nature, and this may cause an imbalance in sentiment analysis and opinion-mining tasks without subjectivity detection.

Pang and Lee [568]; Das and Sagnika [686] applied subjectivity detection to filter out objective sentences in reviews prior to classifying their polarity. Similarly, [688] first extracted subjective sentences from customer reviews and then employed a rule-based system to mine feature-opinion pairs from the subjective sentences. Barbosa and Feng [590]; Soong et al. [689] used subjectivity detection in sentiment analysis for Twitter microtext. These works proved that removing objective content from the dataset indeed makes the learning of sentiment more effective.

6.8.2. Information retrieval

Subjectivity detection can serve as a subsystem in an information retrieval system to determine whether a document is subjective or objective [689], because information retrieval systems normally aim to retrieve either opinionated or factual topic-relevant text from web sources, e.g., tweets, blog posts, reviews webpages, etc. [690].

For opinion retrieval, it helps to select candidate opinionated documents. For instance, [572] first employed an SVM classifier that used unigram and bigram features to identify subjective documents. Then, they separated relevant documents from irrelevant ones. For factual information retrieval, on the other hand, subjectivity detection helps to filter out opinionated text such as allegations and speculations to prevent false hits. Wiebe and Riloff [573] implemented a Naïve Bayes subjectivity classifier and a domain-relevant indicator for selective subjective sentence filtering. If a sentence was classified as subjective, it would be discarded unless it was also labeled as relevant by the indicator.

6.8.3. Hate speech detection

Hate speech detection is a task that identifies abusive speech targeting a person or a group based on stereotypical group characteristics, e.g., ethnicity, religion, or gender, on social media [691]. Since hate speech is often marked by its content, tone, and target [692], its detection is similar to that of polarity. Additionally, subjectivity clues tend to be surrounding the polarizing and arguing topics, which aligns well with hate speech detection. As such, subjectivity detection can be used as a filtering subsystem in hate speech detection.

For instance, [693] employed a rule-based subjectivity classifier that leveraged lexicons including MPQA Subjectivity Lexicon and SentiWordNet to identify subjective sentences. From the extracted sentences, they built a hate speech lexicon using bootstrapping and WordNet. Experiments showed that the addition of subjectivity detection significantly improved the performance of the hate speech classifier.

6.8.4. Question answering system

QA systems generally encounter two types of questions — the ones that expect truth as answers, and the ones that expect opinions. Therefore, it is crucial for a QA system to distinguish opinions from facts, and provide the appropriate type depending on the question [577].

To achieve this goal, a QA system should operate in two stages. First, it must determine whether a question calls for a subjective or objective answer, which is its subjectivity orientation [578,694,695]. Then, the system needs to consider subjectivity as a relevant factor in the information retrieval process.

Subjectivity detection can be incorporated as a filter or feature set in a QA system. For instance, [696] modified the conventional QA system by applying a subjectivity filter and an opinion source filter on the initial IR results, which improved the system significantly. On the other hand, [697] leveraged subjective features from reviews to provide users with a list of relevance-ranked reviews, which improved the performance of answering binary questions from categories with abundant data.

6.9. Summary

Subjectivity detection is a cognitive semantic processing task. It categorizes statements by subjective and objective classes. Theoretical research indicates that subjectivity can be detected by certain subjective elements, e.g., morphological, lexical, and syntactic elements [579]. Thus, computational subjectivity research has developed lexical resources, e.g., Subjectivity Clues [562], and MPQA Subjectivity Lexicon [604]. On the other hand, subjectivity can be also explained from the perspectives of pragmatics, e.g., speech acts [582] and conceptual metaphors [539]. Related subjectivity detection works defined the task as classification tasks. Although those classification tasks can be further divided into course-grained and fine-grained classifications, e.g., document-level, sentence-level, and span-level subjectivity detection, there have not been studies aimed at explaining the subjectivity from pragmatic perspectives, e.g., speech acts and metaphors.

The application of subjectivity detection has proven to be supportive in downstream tasks, such as sentiment computing, information retrieval, hate speech detection, and QA systems. This is because these downstream tasks normally aim at mining opinions from subjective expressions. Subjectivity detection can filter out the objective ones, thus yielding the desired input for downstream tasks.

Table 21
A summary of representative subjectivity detection techniques (Part 1). ISD denotes individual subjectivity detection. SCSL denotes self-collected subjectivity lexicon. SWSD denotes subjectivity WSD.

Task	Reference	Techniques	Feature and KB	Framework	Dataset	Score	Metric
ISD	Riloff and Wiebe [562]	Rule	SCSL	Logic Rules	–	–	–
	Kim and Hovy [563]	Statistics	MPQA, WN	WN distance	MPQA	65.00%	Acc
	Benamara et al. [608]	Statistics	Lexical, stylistic, syntactic, discursive features	SVM	Self-collected	82.31%	Acc
	Xuan et al. [567]	Statistics	MPQA, syntax-based patterns	MaxEnt	Movie	92.10%	Acc
	Remus [610]	Statistics	MPQA, readability	SVM	Movie	84.50%	F1
	Barbosa and Feng [590]	Statistics	MPQA, POS,tweet-specific features	SVM	Twitter1	81.59%	Acc
	Sixto et al. [613]	Statistics	MPQA, tweet-specific features	Stacking classifier	TASS	89.80%	Acc
	Keshavarz and Saniee Abadeh [615]	Statistics	SCSL	Genetic algorithm	SemEval2013	60.90%	Acc
	Khatua et al. [592]	DL	SenticNet	CNN	NET	80.70%	Acc
	Akkaya et al. [616]	Statistics	MPQA, SWSD	SVM	MPQA	81.30%	Acc
	Ortega et al. [620]	Rule	MPQA, SWSD	Clustering, logic rules	Movie	55.68%	F1
	Kamil et al. [571]	Statistics	Pruned ICF	ANFIS	Movie	91.66%	Acc
	Rustamov [624]	Statistics	Pruned ICF	FCS, ANFIS, HMM	Movie	92.24%	Acc
	Wang and Manning [625]	Statistics	Word presence	NBSVM	Movie	93.60%	Acc
	Maas et al. [629]	Statistics	Semantic and sentiment embeddings	Probabilistic model, LDA .	MPQA	71.20%	Acc
	Lin et al. [630]	Statistics	Sentiment	LDA	Movie	88.58%	Acc
	Zhao et al. [632]	DL	Word2Vec	CNN	Movie	95.50%	Acc
	Chaturvedi et al. [634]	DL	MPQA, POS	ELM, RNN, CNN, fuzzy classifier	MPQA Gold	75.00%	Acc
	Satapathy et al. [593]	DL	GloVe, MPQA	CNN, PPR	MPQA	50.00%	F1
	Al Hamoud et al. [576]	DL	GloVe	RNN, Att	Movie	92.80%	Acc
	Kim [637]	DL	Word2Vec	CNN	Movie	93.40%	Acc
	Huo and Iwaihara [641]	DL	BERT	MTL	Movie	95.23%	Acc
	Satapathy et al. [642]	DL	BERT	MTL, RNN,NTN	Movie	95.10%	Acc
	Sagnika et al. [646]	DL	Sentiment-enhanced word embedding	CNN, LSTM	Movie	97.10%	Acc

6.9.1. Technical trends

Subjectivity detection is a well-studied sub-problem in affective computing and opinion mining. There are five technical trends in this area, namely individual, context-dependent, cross-lingual, multi-modal subjectivity detection, and bias detection. A summary of the trends can be found in Tables 21 and 22.

For individual subjectivity detection (Table 21), the subjectivity of each sentence or snippet is determined only by the lexical, syntactic, and semantic information of the sentence itself. There are mainly three types of methods for individual subjectivity detection. First, the lexicon-based approaches rely on external lexicons that contain subjective and sentiment clues to predict the subjectivity of a sentence. The weakness of such an approach is that subjective clues are often not extensive and reliable enough to determine the subjectivity of a sentence. Some works attempted to address this issue by utilizing sentence-level features to extract syntactic information [566,567,590], or incorporating WSD to identify subjective clues according to context [616,620]. Nonetheless, these methods cannot fully extract the underlying sentence structure and contextual information. Word-frequency-based approaches, on the other hand, predict sentence subjectivity according to the word presence or occurrence in a given corpus, thus being able to adapt to new domains and languages. Additionally, this approach requires little external resources or human effort. However, similar to the lexicon-based approach, word frequency methods lack the ability to capture syntactic information. To address this limitation, deep-learning-based methods utilize neural networks to learn spatial and temporal dependencies. Specifically, PLMs are widely used for their ability to provide universal representations [637–639].

For context-dependent subjectivity detection (Table 22), the subjectivity of a sentence is determined with regards to its surrounding context, e.g., inter-sentence-level [568,656], document-level [577, 649,654], or discourse-level [569] information. In the existing works, such information is typically captured through feature engineering or statistical means.

As a large part of subjectivity detection works to some extent relies on external subjective clues, cross-lingual subjectivity detection aims specifically to solve the lack of lexical resources for non-English languages. There are mainly two branches of thought to address this problem (Table 22). One is to make use of language-independent methods such as word frequency [570,571,630,656] and language modeling [654]. The other is to generate resources for the target language from English lexicons with the help of SMT systems [588, 666]. Multi-modal subjectivity detection is a rising field of interest in accordance with the rising need for sentiment analysis in various media (Table 22). Existing works utilized lexical, prosodic, and phonemic features for subjectivity detection in spoken conversations [596,673]. Subjectivity detection in other modalities such as video remains mostly unexplored. Bias detection is a task that is closely related to subjectivity detection (Table 22). It aims to identify biased statements from supposedly impartial articles such as Wikipedia. Despite its greater complexity, the identification of bias exhibits technical patterns that are akin to those found in subjectivity detection, e.g., lexicon-based [601, 678], deep learning [602,683], and cross-lingual [680] methods.

6.9.2. Application trends

Due to its filtering nature, subjectivity detection is widely used as a parser for many downstream tasks, e.g., sentiment analysis [568, 590,686,688,689], information retrieval [572,573], hate speech detection [693], and QA systems [696,697]. Most existing works take the pipeline approach, using the filtered results from subjectivity detection as the input of the target application. On the other hand, we also observe that subjectivity lexicons can also be useful features to support hate speech detection and QA systems (see Table 23). A survey of literature pertaining to subjectivity detection reveals that the progress made in this research area has not kept pace with the advancements made in its downstream sentiment computing tasks, e.g., sentiment analysis [698]. This is likely because sentiment analysis may deliver

Table 22
A summary of representative subjectivity detection techniques (Part 2). CDSD denotes concept-dependent subjectivity detection. CLSD denotes cross-lingual subjectivity detection. MMSD denotes multi-modal subjectivity detection. BD denotes bias detection. SWN denotes SentiWordNet.

Task	Reference	Techniques	Feature and KB	Framework	Dataset	Score	Metric
CDSD	Pang and Lee [568]	Statistics	SCSL	Minimum cuts, Naïve Bayes	Movie	86.40%	Acc
	Das and Bandyopadhyay [649]	Rule	MPQA, doc-level features, SWN	Logic rules	MPQA	79.54%	F1
	Das and Bandyopadhyay [650]	Statistics	MPQA, POS, doc-level features	Genetic algorithm	Movie	95.69%	F1
	Biyani et al. [569]	Statistics	MPQA, SentiStrength, thread-specific features	Logistic regression	Forum	77.01%	Acc
	Yu and Hatzivassiloglou [577]	Statistics	MPQA, POS	Naïve Bayes	TREC	97.00%	F1
	Karimi and Shakery [654]	Statistics	Language model	Rank by similarity	Movie	94.63%	MAP
	Chaturvedi et al. [657]	DL	MPQA	GBN, CNN	Movie	96.40%	Acc
CLSD	Banea et al. [588]	ML	MPQA	SMT, Naïve Bayes	Multi-MPQA EN	74.72%	Acc
	Mogadala and Varma [659]	Statistics	Unigram and bigram freq., word length	Naïve Bayes	Multi-MPQA EN	92.50%	F1
	Chaturvedi et al. [666]	DL	MPQA, WSD	SMT, CNN, RNN	MPQA Gold	84.00%	F1
MMSD	Murray and Carenini [596]	Statistics	VIN, raw features	MaxEnt	AMIDA	52.00%	F1
	Raaijmakers et al. [673]	Statistics	Lexical, prosodic, and phonemic features	BoosTexter	AMIDA	75.40%	Acc
BD	Recasens et al. [678]	Statistics	Biased lexicon, POS	Logistic regression	Self-collected	34.35%	Acc
	Hube and Fetahu [601]	Statistics	Word2Vec, POS, LIWC, biased lexicon	Random Forest	Conservapedia	74.00%	Prec.
	Aleksandrova et al. [680]	Statistics	Word frequency	BoW	Self-collected	59.57%	F1
	Hube and Fetahu [683]	DL	GloVe, POS, LIWC	RNN	Self-collected	77.10%	F1
	Pryzant et al. [602]	DL	BERT	LSTM	WNC	45.80%	Acc
	Pant et al. [685]	DL	BERT	Ensemble, BERT	WNC	71.61%	Acc

Table 23
A summary of the representative applications of subjectivity detection in downstream tasks.

Reference	Downstream tasks	Feature	Parser
Bonzanini et al. [687]	Sentiment computing		✓
Pang and Lee [568]	Sentiment computing		✓
Das and Sagnika [686]	Sentiment computing		✓
Kamal [688]	Sentiment computing		✓
Barbosa and Feng [590]	Sentiment computing		✓
Soong et al. [689]	Sentiment computing		✓
Zhang et al. [572]	Information retrieval		✓
Wiebe and Riloff [573]	Information retrieval		✓
Cohen-Almagor [692]	Hate speech detection		✓
Gitari et al. [693]	Hate speech detection	✓	✓
Li et al. [578]	Question answering	✓	✓
Li et al. [694]	Question answering	✓	✓
Aikawa et al. [695]	Question answering	✓	✓
Stoyanov et al. [696]	Question answering		✓
Wan and McAuley [697]	Question answering	✓	

more fine-grained classification outputs, which helps to gain business insights, e.g., sentiment polarities on product or service reviews. However, it should be noted that while positive, negative, and neutral sentiment polarities represent subsets of subjective texts, there exists a substantial portion of texts that are objective in nature, presenting factual information. Objective texts are likely to be infrequent in reviews of products or services, as customers often use such platforms to express their opinions. However, in the context of opinion mining on social media, it is crucial to differentiate between subjective and objective statements, given that even statements with neutral sentiment polarities can be indicative of an individual's opinion. Thus, it is still necessary to conduct subjectivity detection before sentiment analysis.

6.9.3. Future works

Fine-grained subjectivity detection. A sentence may contain several clauses with differing subjectivity. For instance, a sentence may present two or more opinions, or contain both opinions and factual information. Therefore, to better assist downstream applications, fine-grained subjectivity detection that identifies the particular opinion-bearing clauses is worthy of investigation. However, there is limited research on this issue. Benamara et al. [608] proposed segment-level subjectivity detection. Wilson et al. [566] proposed a method specifically for classifying the subjectivity of deeply nested clauses. There is scope for additional research to exploit the full potential of the fine-grained subjectivity annotation offered by the MPQA scheme [584].

Multi-modal subjectivity detection. Subjectivity detection using information from multiple modalities remains largely unexplored. There is related multi-modal research that might provide inspiration for future works. Wrede and Shriberg [674] aimed to identify hot spots, which are regions in a meeting where participants are highly involved in the discussion, using solely a set of prosodic features. Hillard et al. [699]; Galley et al. [700] both targeted the detection of agreements and disagreements in meetings. The former explored the combination of lexical and prosodic features, whereas the latter incorporated pragmatic features that captured the interactions between speakers. Neiberg et al. [701] recognized positive, negative, and neutral emotions in meetings using lexical and acoustic-prosodic features. Somasundaran et al. [702] detected sentences and turns in meetings that express sentiment and arguing opinions using lexical and discourse features. Morency et al. [600]; Wöllmer et al. [599]; Tsai et al. [703] conducted sentiment analysis on review videos using linguistic features, acoustic features, and visual features (face tracking).

Explainable subjectivity detection. While much of the subjectivity detection research has utilized lexical resources such as subjectivity and affective lexicons to explain the subjective nature of text based on individual words, these resources do not capture the pragmatic nuances of words within their contextual environment. This is because the utilized lexical knowledge is context-independent. Theoretical research has explained subjectivity from the perspective of pragmatics [539,582]. It would be valuable to study subjectivity detection that detects and explains subjectivity. Explainable subjectivity detection could push the development of more linguistics-inspired models that can account for the complexities of subjectivity and its expression in natural language. Additionally, there is potential for cross-disciplinary collaboration between linguistics, cognitive science, and computer science to further advance our understanding of subjectivity and its detection in various domains.

7. Discussion

7.1. Interactions between the surveyed tasks

In preceding sections, we have provided an introduction to the relationships between our surveyed tasks and downstream applications. Nevertheless, it is important to recognize that these tasks are intrinsically interconnected. For example, WSD and anaphora resolution are mutually supportive for each other. Consider the following sentence:

(21) I observed a colossal mammoth statue on the summit. It’s really cool.

In this case, an anaphora resolution model should be capable of linking “it” to the “mammoth statue”, assuming the significance of “cool” is interpreted as “a form of approval due to the appealing attributes of the mammoth statue”, rather than the low temperature associated with the “summit”. Conversely, if the antecedent of “it”, denoting the “mammoth statue”, is established, the intended meaning of “cool” can be easily discerned. This symbiotic enhancement is also observable in the context of WSD and NER. Within the context of the following sentence, disambiguating the sense of “hit” aids NER in recognizing “King’s Arm” as a location rather than a person.

(22) I hit the King’s Arm yesterday. It’s my preferred pub in London.

Likewise, accurately identifying “King’s Arm” as a location bolsters WSD models in determining that “hit” should be interpreted as “visited”.

In the domain of concept extraction, WSD for multi-word expressions assumes heightened significance. For example, “go bananas”, “cloud computing”, and “pain killer” are best captured as concepts with multi-word expressions, rather than independent words, since their meanings manifest coherently only when interpreted as integrated wholes. Absent WSD for multi-word expressions, the task of concept extraction struggles to delineate conceptual boundaries within a given sentence. Furthermore, the application of WSD techniques extends to textual subjectivity detection. Taking the adjective “fine” for example, it ordinarily corresponds with the subjective text due to its meaning referring to the subjective feeling of being satisfactory, as seen in “Tesla Model X is a fine car”. However, “fine” can also appear in objective contexts, if construed as a monetary penalty, as demonstrated in “I received a fine yesterday for speeding”. Another instance is the term “long”, which can be employed in an objectively spatial context as well as a negatively subjective sense akin to “tenacious”. Integrating a sense-sensitive approach into subjectivity detection can enhance its performance.

The aforementioned interconnectedness and instances highlight the intricate interplay of language. The explication of linguistic interpretations can encompass various dimensions, even though the surveyed tasks pertain to fundamental semantic endeavors. These tasks exhibit interdependencies and mutual dependencies. Consequently, diverse learning methods may be requisite for addressing these multidimensional linguistic interpretation tasks.

7.2. The impacts of deep learning on semantic processing

In the current neural network models with end-to-end task-proces sing purposes, the aforementioned linguistic interpretation facets might be encapsulated within a black box, lacking explicit representation. The limitation of these approaches lies in their inability to elucidate how language is employed and construed across divergent semantic facets. While the pursuit of human-like accuracy in deep learning-based systems is prominent, it is essential to acknowledge that the simulation of human cognitive and interpretive mechanisms, akin to human-like intelligence, may just gain a secondary focus in contrast to the endeavor for heightened task accuracy in the NLP domain.

Prior to the era of deep learning, semantic processing tasks often relied on rule-based or symbolic methods [704]. These approaches aimed to distill the linguistic intuitions and insights associated with a given semantic processing task by leveraging a variety of linguistic features. Algorithms were devised to capture the specific linguistic nuances of each task, and substantial endeavors were undertaken to unveil the overarching principles governing semantic interpretation [535,616]. The process of cross-validating different linguistic features played a prominent role in the pursuit of enhanced predictive accuracy.

Nonetheless, the emergence of neural networks has brought about a convergence in the landscape of semantic learning and representations. Within the domain of semantic learning, a prevalent strategy involves utilizing contextual cues to predict a target word. This is achieved through various learning paradigms such as continuous bag of words [word2vec, 23], masked word prediction [as seen in models like BERT and RoBERTa, 8,9], or the prediction of the subsequent word [exemplified by the GPT families, 705–707]. This approach has undoubtedly yielded remarkable accomplishments across a wide range of NLP tasks. Neural networks excel at capturing the fundamental meanings of words and sentences within vectorized representations, and their ability to encode contextualized meanings as the network architecture becomes deeper.

In light of the demonstrated efficacy of the aforementioned neural semantic learning paradigms, the emphasis on tailoring models to capture task-specific linguistic intuitions has diminished somewhat, compared to rule-based or symbolic methods. Nevertheless, a pertinent query arises: Is the general unified target word prediction approach of pre-training the optimal strategy for achieving multi-dimensional semantic understanding? Semantic representations in vector form possess the capacity to apprehend spatial correlations among meanings, manifesting through distinct proximities of similar and dissimilar meanings. These spatial relationships are forged through the learning of word associations. Nonetheless, substantial knowledge, such as commonsense, causality, and occurrences that are either unprecedented or infrequent, e.g., novel metaphors [708], remain beyond the direct purview of contextual understanding. Consequently, the comprehension of intricate constructs like frame semantics [28], narratives, and cognitive mechanisms – which intricately hinge on facets like knowledge representation, commonsense reasoning, social cognition, and learning – presents challenges when solely relying on vector representations for their explication [709].

Considering the strong connections between semantic processing tasks and linguistics, it is advisable to direct heightened attention toward the incorporation of linguistic and cognitive intuitions and the exploration of semantic interpretative dimensions via neural networks and neuro-symbolic methods that marry the advantages of neural nets and symbolic knowledge representations. This constitutes a salient characteristic demarcating computational semantics-focused research from pure machine learning-oriented deep learning studies, inspiring a broader exploration of semantic processing.

7.3. Semantic processing and large language models

ChatGPT and GPT-4 have expanded the reach of LLMs across diverse domains. Their remarkable proficiency in text generation, multitask execution, and complex task handling has garnered significant attention within the NLP community. Meantime, it is evident that there are noteworthy challenges associated with these expansive models, including issues such as hallucinations and complex task reasoning [710]. In the context of dialogues, [711] have introduced an innovative evaluation framework tailored for LLMs. Their study has compared multiple such models and identified a recurring challenge of hallucination – a scenario where the generated content appears plausible but is, in fact, entirely fictional. Embedding semantic knowledge into these models presents an avenue to offer them with a more precise comprehension of real-world information, thereby diminishing the likelihood of generating content that lacks substantiated foundation.

For instance, consider the sentence "the horse flew over the barn". A model enriched with semantic acumen would promptly discern the implausibility of such an event, thereby reducing the susceptibility to produce hallucinatory output. The scope of semantic acumen encompasses not solely the literal meaning of "barn", but also encompasses a more widespread understanding of its prevalent dimensions of size and height. Such a semantically informed model can manifest as a system adept at recognizing inconsistencies or deviations from anticipated semantic patterns. Alternatively, it could be a model proficient in grasping ordinary concept associations grounded in the frame semantics. Moreover, semantic processing can assist in reformulating user queries to render them more machine-friendly, thereby mitigating the potential for hallucinations stemming from vague queries. On the other hand, [710] have reported that ChatGPT demonstrates satisfactory performance in general scientific knowledge and can effectively address questions necessitating open-ended responses. Nonetheless, it is not without errors, particularly in cases requiring multi-step reasoning. This shortcoming may be attributed to the current practice of employing solely feedforward propagation and fast inference in LLMs [712]. The absence of human-like deliberation for complex inquiries impedes the model's capacity for intricate multi-step reasoning tasks. Recent strides in Chain-of-Thought Prompting [713] underscore the potential of decomposing complex problems into intermediate steps to enhance the complex reasoning capabilities of LLMs. In this context, semantic processing emerges as a valuable asset for task decomposition. It can assist in identifying pivotal concepts and entities, and delineating the principal topic into coherent logical subtopics or sequential steps. Semantic comprehension ensures a seamless and coherent progression of steps, yielding prompts that are not only more efficient but also conducive to the adept reasoning of intricate challenges by LLMs.

8. Conclusion

In this survey, we have reviewed recent semantic processing techniques, e.g., WSD, anaphora resolution, concept extraction, NER, and subjectivity detection. We summarized useful datasets, annotation tools and knowledge bases that can facilitate the research in these domains. We also summarized the technical trends of these techniques, related theoretical research, and their downstream applications. We found that the breadth and depth of semantic processing can be greatly extended, both from the perspective of the needs of theoretical research and downstream applications. This is because current computational semantic processing techniques are limited in their reliance on specific task settings and available datasets. The review of the downstream applications of semantic processing techniques could potentially stimulate further research into fusion methodologies, which seek to enhance the performance of downstream tasks. The semantic processing methods can not only deliver effective features for downstream tasks, but also gain insights into analyzing model behaviors and studying linguistic and cognitive patterns.

As we continue to advance in the field of NLP, using powerful PLMs and LLMs has become increasingly common to tackle more complex NLP tasks. However, it is important to note that there is still great academic value in studying the low-level semantic tasks that these models are built upon. These tasks help us understand how language is presented and received, how semantics relates to human cognition, and how semantic processing tasks are interrelated. We observe that numerous contemporary semantic processing tasks have been translated into machine learning problems, which have somehow diminished linguistic motivations and intuitions from these computational studies. Shaping semantic processing tasks into tasks that are more conducive to machine learning can indeed improve the accuracy of specific tasks. However, improving accuracy in a single-task setting is not the only pursuit of semantic processing. We should pay more attention to how semantic processing techniques can better serve humans and machines to explain language phenomena.

We hope that this paper can stimulate more research directions in the field of semantic processing and inspire researchers to place greater emphasis on the nature and cognition of semantics. With the development of more powerful tools such as PLMs and LLMs, it is perhaps valuable for our research to use these tools to address those fundamental linguistic challenges that were previously considered daunting. Regardless of the sophistication of tasks that can be performed by LLMs, basic semantic processing tasks remain crucial for comprehending and utilizing language effectively. These tasks serve as the foundation upon which our understanding of language is built.

CRediT authorship contribution statement

Rui Mao: Conceptualization, Methodology, Investigation, Writing – original draft (Introduction, Discussion, and Conclusion), Harmonization. **Kai He:** Investigation, Writing – original draft (Named Entity Recognition). **Xulang Zhang:** Investigation, Writing – original draft (Subjectivity Detection). **Guanyi Chen:** Investigation, Writing – original draft (Anaphora Resolution). **Jinjie Ni:** Investigation, Writing – original draft (Word Sense Disambiguation). **Zonglin Yang:** Investigation, Writing – original draft (Concept Extraction). **Erik Cambria:** Conceptualization, Writing – review & editing, Project administration, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Erik Cambria reports financial support was provided by Continental Automotive Singapore Pte Ltd. Rui Mao reports financial support was provided by Continental Automotive Singapore Pte Ltd. Zonglin Yang reports financial support was provided by Continental Automotive Singapore Pte Ltd.

Data availability

No data was used for the research described in the article.

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] X. Zhang, R. Mao, E. Cambria, A survey on syntactic processing techniques, Artif. Intell. Rev. 56 (2023) 5645–5728, http://dx.doi.org/10.1007/s10462-022-10300-7.

[2] F.R. Palmer, P. Frank Robert, Semantics, Cambridge University Press, 1981.

[3] G.E. Noyes, The first English dictionary, Cawdrey's table alphabeticall, Mod. Lang. Notes 58 (8) (1943) 600–605.

[4] J. Simpson, E. Weiner, The Oxford English Dictionary, second ed., Oxford University Press, 1989.

[5] W. Croft, D.A. Cruse, Cognitive Linguistics, Cambridge University Press, 2004.

[6] J. Barwise, J. Perry, Situations and attitudes, J. Philos. 78 (11) (1981) 668–691.

[7] R. Jackendoff, Toward an explanatory semantic representation, Linguist. Inquiry 7 (1) (1976) 89–150.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional Transformers for language understanding, in: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, arXiv e-prints, arXiv–1907.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[11] S.A. Salloum, R. Khan, K. Shaalan, A survey of semantic analysis approaches, in: Proceedings of the International Conference on Artificial Intelligence and Computer Vision, AICV2020, Springer, 2020, pp. 61–70.

[12] R. Ransing, A. Gulati, A survey of different approaches for word sense disambiguation, in: ICT Analysis and Applications: Proceedings of ICT4SD 2022, Springer, 2022, pp. 435–445.

[13] M. Poesio, J. Yu, S. Paun, A. Aloraini, P. Lu, J. Haber, D. Cokal, Computational models of Anaphora, Ann. Rev. Linguist. 9 (2023) 561–587.

[14] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K.J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, et al., Clinical concept extraction: A methodology review, J. Biomed. Inform. 109 (2020) 103526.

[15] Y. Wang, H. Tong, Z. Zhu, Y. Li, Nested named entity recognition: A survey, ACM Trans. Knowl. Discov. Data 16 (6) (2022) 1–29.

[16] A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments, Decis. Support Syst. 53 (4) (2012) 675–679.

[17] R. Navigli, S.P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250.

[18] M. Wang, Y. Wang, A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 6229–6240.

[19] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4330–4338, http://dx.doi.org/10.24963/ijcai.2021/593.

[20] R. Navigli, Word sense disambiguation: A survey, ACM Comput. Surv. (CSUR) 41 (2) (2009) 1–69.

[21] J. Firth, A synopsis of linguistic theory, 1930–1955, Stud. Linguist. Anal. (1957) 10–32.

[22] P.D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, J. Artificial Intelligence Res. 37 (2010) 141–188.

[23] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).

[24] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inform. Process. Syst. 30 (2017).

[26] Y. Wilks, Preference semantics, Computer Science Department, Stanford University, 1973.

[27] A. Goldberg, L. Suttle, Construction grammar, Wiley Interdisc. Rev.: Cogn. Sci. 1 (4) (2010) 468–477.

[28] C.J. Fillmore, et al., Frame semantics, Cogn. Linguist.: Basic Readings 34 (2006) 373–400.

[29] M.R. Petruck, Frame semantics, Handb. Pragmat. 2 (1996).

[30] J. Ruppenhofer, M. Ellsworth, M. Schwarzer-Petruck, C.R. Johnson, J. Scheffczyk, FrameNet II: Extended Theory and Practice, Technical Report, International Computer Science Institute, 2016.

[31] G.A. Miller, C. Leacock, R. Tengi, R.T. Bunker, A semantic concordance, in: Human Language Technology: Proceedings of a Workshop Held At Plainsboro, New Jersey, March 21-24, 1993, 1993, pp. 303–308.

[32] E. Pianta, L. Bentivogli, C. Girardi, MultiWordNet: Developing an aligned multilingual database, in: First International Conference on Global WordNet, 2002, pp. 293–302.

[33] C. Leacock, G. Towell, E.M. Voorhees, Corpus-based statistical sense resolution, in: Human Language Technology: Proceedings of a Workshop Held At Plainsboro, New Jersey, March 21-24, 1993, 1993, pp. 260–265.

[34] R. Bruce, J. Wiebe, Decomposable modeling in natural language processing, Comput. Linguist. 25 (2) (1999) 195–207.

[35] H.T. Ng, H.B. Lee, Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach, in: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, 1996, pp. 40–47.

[36] T. Chklovski, P. Pantel, Verbocean: Mining the web for fine-grained semantic verb relations, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 33–40.

[37] K. Taghipour, H.T. Ng, One million sense-tagged instances for word sense disambiguation and induction, in: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, 2015, pp. 338–344.

[38] P. Edmonds, S. Cotton, SensEval-2: Overview, in: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, 2001, pp. 1–5.

[39] B. Snyder, M. Palmer, The English all-words task, in: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004, pp. 41–43.

[40] S. Pradhan, E. Loper, D. Dligach, M. Palmer, SemEval-2007 task-17: English lexical sample, SRL and all words, in: Proceedings of the Fourth International Workshop on Semantic Evaluations, SemEval-2007, 2007, pp. 87–92.

[41] R. Navigli, D. Jurgens, D. Vannella, SemEval-2013 task 12: Multilingual word sense disambiguation, in: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval 2013, 2013, pp. 222–231.

[42] A. Moro, R. Navigli, SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking, in: Proceedings of the 9th International Workshop on Semantic Evaluation, 2015, pp. 288–297.

[43] L. Vial, B. Lecouteux, D. Schwab, Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation, in: Proceedings of the 10th Global WordNet Conference, 2019, pp. 108–117.

[44] L. Bentivogli, E. Pianta, Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus, Nat. Lang. Eng. 11 (3) (2005) 247–261.

[45] F. Bond, T. Baldwin, R. Fothergill, K. Uchimoto, Japanese SemCor: A sense-tagged corpus of Japanese, in: Proceedings of the 6th Global WordNet Conference, 2012, pp. 56–63.

[46] T. Pasini, R. Navigli, Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 78–88.

[47] T. Pasini, F. Elia, R. Navigli, Huge automatically extracted training-sets for multilingual word sense disambiguation, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018, pp. 1694–1698.

[48] B. Scarlini, T. Pasini, R. Navigli, Just “OneSeC” for producing multilingual sense-annotated data, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 699–709.

[49] T. Pasini, R. Navigli, Train-o-matic: Supervised word sense disambiguation with no (manual) effort, Artificial Intelligence 279 (2020) 103215.

[50] B.T. Atkins, Tools for computer-aided corpus lexicography: The hector project, Acta Linguist. Hung. 41 (1/4) (1992) 5–71.

[51] A. Raganato, J. Camacho-Collados, R. Navigli, Word sense disambiguation: A unified evaluation framework and empirical comparison, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 99–110.

[52] M. Mayor, Longman Dictionary of Contemporary English, Pearson Education India, 2009.

[53] O. Dictionary, Oxford dictionary of english, in: Oxford Dictionary of English, third ed., Oxford University Press. China Translation & Printing Services Ltd, China, 2010.

[54] C.E. Dictionary, Collins, Lond. Glasg. (1982).

[55] A.S. Hornby, A.P. Cowie, Oxford Advanced Learner’s Dictionary of Current English, Paperback, 1974.

[56] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller, Introduction to WordNet: An on-line lexical database, Int. J. Lexicogr. 3 (4) (1990) 235–244.

[57] S. Wang, F. Bond, Building the Chinese open wordnet (COW): Starting from core synsets, in: Proceedings of the 11th Workshop on Asian Language Resources, 2013, pp. 10–18.

[58] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, C. Fellbaum, et al., Introducing the Arabic WordNet project, in: Proceedings of the Third International WordNet Conference, 2006, pp. 295–300.

[59] M. Postma, E. Van Miltenburg, R. Segers, A. Schoen, P. Vossen, Open Dutch WordNet, in: Proceedings of the 8th Global WordNet Conference, GWC, 2016, pp. 302–310.

[60] M. Maru, F. Scozzafava, F. Martelli, R. Navigli, SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 3534–3540.

[61] P. Resnik, D. Yarowsky, Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation, Nat. Lang. Eng. 5 (2) (1999) 113–133.

[62] T. Cohn, Performance metrics for word sense disambiguation, in: Proceedings of the Australasian Language Technology Workshop 2003, 2003, pp. 86–93.

[63] S. Neale, J. Silva, A. Branco, A flexible tool for manual word sense annotation, in: Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, ISA-11, 2015, pp. 1–5.

[64] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, in: Proceedings of the 5th Annual International Conference on Systems Documentation, 1986, pp. 24–26.

[65] S. Banerjee, T. Pedersen, et al., Extended gloss overlaps as a measure of semantic relatedness, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Vol. 3, 2003, pp. 805–810.

[66] P. Basile, A. Caputo, G. Semeraro, An enhanced lesk word sense disambiguation algorithm through a distributional semantic model, in: Proceedings of the 25th International Conference on Computational Linguistics, 2014, pp. 1591–1600.

[67] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[68] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.

[69] E. Agirre, O. López de Lacalle, A. Soroa, Random walks for knowledge-based word sense disambiguation, Comput. Linguist. 40 (1) (2014) 57–84.

[70] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: A unified approach, Trans. Assoc. Comput. Linguist. 2 (2014) 231–244.

[71] R. Tripodi, R. Navigli, Game theory meets embeddings: a unified framework for word sense disambiguation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 88–99.

[72] E. Agirre, A. Soroa, Personalizing pagerank for word sense disambiguation, in: Proceedings of the 12th Conference of the European Chapter of the ACL, 2009, pp. 33–41.

[73] T.H. Haveliwala, Topic-sensitive PageRank, in: Proceedings of the 11th International Conference on World Wide Web, 2002, pp. 517–526.

[74] R. Navigli, M. Lapata, Graph connectivity measures for unsupervised word sense disambiguation, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007, pp. 1683–1688.

[75] H. Tong, C. Faloutsos, J.-Y. Pan, Fast random walk with restart and its applications, in: Sixth International Conference on Data Mining, IEEE, 2006, pp. 613–622.

[76] F. Scozzafava, M. Maru, F. Brignone, G. Torrisi, R. Navigli, Personalized PageRank with syntagmatic information for multilingual word sense disambiguation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 37–46.

[77] M. Kågebäck, H. Salomonsson, Word sense disambiguation using a bidirectional LSTM, in: Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon, CogALex-V, 2016, pp. 51–56.

[78] S. Kumar, S. Jat, K. Saxena, P. Talukdar, Zero-shot word sense disambiguation using sense definition embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5670–5681.

[79] T. Blevins, L. Zettlemoyer, Moving down the long tail of word sense disambiguation with gloss informed Bi-encoders, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1006–1017.

[80] M. Bevilacqua, R. Navigli, Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2854–2864.

[81] S. Conia, R. Navigli, Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021, pp. 3269–3275.

[82] E. Barba, L. Procopio, C. Lacerra, T. Pasini, R. Navigli, Exemplification modeling: Can you give me an example, please? in: IJCAI, 2021, pp. 3779–3785.

[83] R.L. Singh, K. Ghosh, K. Nongmeikapam, S. Bandyopadhyay, A decision tree based word sense disambiguation system in Manipuri language, Adv. Comput. 5 (4) (2014) 17.

[84] T. O'Hara, R. Bruce, J. Donner, J. Wiebe, Class-based collocations for word sense disambiguation, in: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004, pp. 199–202.

[85] Z. Zhong, H.T. Ng, It makes sense: A wide-coverage word sense disambiguation system for free text, in: Proceedings of the ACL 2010 System Demonstrations, 2010, pp. 78–83.

[86] S. Rothe, H. Schütze, AutoExtend: Extending word embeddings to embeddings for synsets and Lexemes, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1793–1803.

[87] I. Iacobacci, M.T. Pilehvar, R. Navigli, Embeddings for word sense disambiguation: An evaluation study, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 897–907.

[88] A. Popov, Word sense disambiguation with recurrent neural networks, in: Proceedings of the Student Research Workshop Associated with RANLP 2017, 2017, pp. 25–34.

[89] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.

[90] D. Yuan, J. Richardson, R. Doherty, C. Evans, E. Altendorf, Semi-supervised word sense disambiguation with neural models, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1374–1385.

[91] P.P. Talukdar, K. Crammer, New regularized algorithms for transductive learning, in: Machine Learning and Knowledge Discovery in Databases: European Conference, Springer, 2009, pp. 442–457.

[92] M. Le, M. Postma, J. Urbani, P. Vossen, A deep dive into word sense disambiguation with LSTM, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 354–365.

[93] C. Hadiwinoto, H.T. Ng, W.C. Gan, Improved word sense disambiguation using pre-trained contextualized word representations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 5297–5306.

[94] M. Bevilacqua, R. Navigli, Quasi bidirectional encoder representations from transformers for word sense disambiguation, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, 2019, pp. 122–131.

[95] A. Raganato, C.D. Bovi, R. Navigli, Neural sequence learning models for word sense disambiguation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1156–1167.

[96] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 933–941.

[97] D. Loureiro, A. Jorge, Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5682–5691.

[98] T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, no. 1, 2018, pp. 1811–1818.

[99] L. Huang, C. Sun, X. Qiu, X.-J. Huang, GlossBERT: BERT for word sense disambiguation with gloss knowledge, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 3509–3514.

[100] B. Scarlini, T. Pasini, R. Navigli, SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 05, 2020, pp. 8758–8765.

[101] B. Scarlini, T. Pasini, R. Navigli, With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 3528–3539.

[102] E. Barba, T. Pasini, R. Navigli, ESC: Redesigning WSD with extractive sense comprehension, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4661–4672.

[103] G. Berend, Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 8498–8508.

[104] U. Farooq, T.P. Dhamala, A. Nongaillard, Y. Ouzrout, M.A. Qadir, A word sense disambiguation method for feature level sentiment analysis, in: 2015 9th International Conference on Software, Knowledge, Information Management and Applications, IEEE, 2015, pp. 1–8.

[105] A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C.L. Ngo, Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment, Expert Syst. Appl. 42 (1) (2015) 306–324.

[106] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 2200–2204.

[107] B. Ohana, B. Tierney, Sentiment classification of reviews using SentiWordNet, Proc. ITT 8 (2009).

[108] H. Saggion, A. Funk, Interpreting SentiWordNet for opinion classification, in: Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp. 1129–1133.

[109] A. Devitt, K. Ahmad, Sentiment polarity identification in financial news: A cohesion-based approach, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 984–991.

[110] C. Hung, H.-K. Lin, Using objective words in SentiWordNet to improve word-of-mouth sentiment classification, IEEE Intell. Syst. 28 (02) (2013) 47–54.

[111] C. Hung, S.-J. Chen, Word sense disambiguation based sentiment lexicons for sentiment classification, Knowl.-Based Syst. 110 (2016) 224–232.

[112] R. Krovetz, W.B. Croft, Lexical ambiguity and information retrieval, ACM Trans. Inform. Syst. (TOIS) 10 (2) (1992) 115–141.

[113] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran, Indexing with WordNet synsets can improve text retrieval, in: Usage of WordNet in Natural Language Processing Systems, 1998, pp. 38–44.

[114] J. Gonzalo, A. Penas, F. Verdejo, Lexical ambiguity and information retrieval revisited, in: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999, pp. 195–202.

[115] C. Stokoe, M.P. Oakes, J. Tait, Word sense disambiguation in information retrieval revisited, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003, pp. 159–166.

[116] M. Sanderson, Word sense disambiguation and information retrieval, in: SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer, 1994, pp. 142–151.

[117] R. Blloshmi, T. Pasini, N. Campolungo, S. Banerjee, R. Navigli, G. Pasi, IR like a SIR: Sense-enhanced information retrieval for multiple languages, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 1030–1041.

[118] S.-B. Kim, H.-C. Seo, H.-C. Rim, Information retrieval using word senses: Root sense tagging approach, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 258–265.

[119] A. Rios Gonzales, L. Mascarell, R. Sennrich, Improving word sense disambiguation in neural machine translation with sense embeddings, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, 2017, pp. 11–19.

[120] N. Campolungo, F. Martelli, F. Saina, R. Navigli, DiBiMT: A novel benchmark for measuring word sense disambiguation biases in machine translation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4331–4352.

[121] A. Raganato, Y. Scherrer, J. Tiedemann, The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, 2019, pp. 470–480.

[122] R. Marvin, P. Koehn, Exploring word sense disambiguation abilities of neural machine translation systems, in: Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, 2018, pp. 125–131.

[123] G. Tang, R. Sennrich, J. Nivre, Encoders help you disambiguate word senses in neural machine translation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 1429–1435.

[124] F. Liu, H. Lu, G. Neubig, Handling homographs in neural machine translation, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1336–1345.

[125] H.T. Ng, Getting serious about word sense disambiguation, in: Tagging Text with Lexical Semantics: Why, What, and how?, 1997, pp. 1–7.

[126] R. Mao, C. Lin, F. Guerin, Word embedding and WordNet based metaphor identification and interpretation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, ACL, Association for Computational Linguistics, 2018, pp. 1222–1231.

[127] R. Mao, X. Li, M. Ge, E. Cambria, MetaPro: A computational metaphor processing model for text pre-processing, Inf. Fusion 86–87 (2022) 30–43.

[128] L. Boroditsky, How language shapes thought, Sci. Am. 304 (2) (2011) 62–65.

[129] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, et al., Recent trends in word sense disambiguation: A survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 4330–4338.

[130] R. Mitkov, The Oxford Handbook of Computational Linguistics, Oxford University Press, 2022.

[131] R. Mitkov, Anaphora Resolution, Routledge, 2014.

[132] T. Reinhart, Coreference and bound anaphora: A restatement of the anaphora questions, Linguist. Philos. (1983) 47–88.

[133] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Inf. Fusion 59 (2020) 139–162.

[134] R. Liu, R. Mao, A.T. Luu, E. Cambria, A brief survey on recent advances in coreference resolution, Artif. Intell. Rev. (2023) 1–43.

[135] A. Garnham, Mental Models and the Interpretation of Anaphora, Psychology Press, 2001.

[136] D. Büring, Binding Theory, Cambridge University Press, 2005.

[137] A.K. Joshi, S. Kuhn, Centered logic: The role of entity centered sentence representation in natural language inferencing, in: Proceedings of the 6th International Joint Conference on Artificial Intelligence-Volume 1, 1979, pp. 435–439.

[138] B.J. Grosz, A.K. Joshi, S. Weinstein, Providing a unified account of definite noun phrases in discourse, in: 21st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 1983, pp. 44–50.

[139] B.J. Grosz, A.K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Comput. Linguist. 21 (2) (1995) 203–225.

[140] A. Kehler, Current theories of centering for pronoun interpretation: A critical evaluation, Comput. linguist. 23 (3) (1997) 467–475.

[141] A. Joshi, R. Prasad, E. Miltsakaki, Anaphora resolution: Centering theory approach, Encyclop. Lang. Linguist. 1 (2006) 223–230.

[142] T. Givón, Topic continuity in discourse: The functional domain of switch reference, Switch Ref. Univers. Gramm. 51 (1983) 82.

[143] W. Chafe, Givenness, contrastiveness, definiteness, subjects, topics, and point of view, in: Subject and Topic, Academic Press, 1976.

[144] J.K. Gundel, N. Hedberg, R. Zacharski, Cognitive status and the form of referring expressions in discourse, Language (1993) 274–307.

[145] S.E. Brennan, Centering attention in discourse, Lang. Cogn. process. 10 (2) (1995) 137–167.

[146] R.J. Stevenson, R.A. Crawley, D. Kleinman, Thematic roles, focus and the representation of events, Lang. Cogn. Proc. 9 (4) (1994) 519–548.

[147] J.E. Arnold, Reference Form and Discourse Patterns, Stanford University, 1998.

[148] C.G. Chambers, R. Smyth, Structural parallelism and discourse coherence: A test of centering theory, J. Memory Lang. 39 (4) (1998) 593–608.

[149] K.E. McCoy, M. Strube, Generating anaphoric expressions: Pronoun or definite description? Relat. Discourse/Dialogue Struct. Ref. (1999).

[150] N. Orita, N. Feldman, J. Boyd-Graber, E. Vornov, Quantifying the role of discourse topicality in speakers' choices of referring expressions, in: Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics, 2014, pp. 63–70.

[151] N. Orita, E. Vornov, N. Feldman, H. Daumé III, Why discourse affects speakers' choice of referring expressions, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Vol. 1, 2015, pp. 1639–1649.

[152] G. Chen, K. van Deemter, C. Lin, Modelling pro-drop with the rational speech acts model, in: Proceedings of the 11th International Conference on Natural Language Generation, Association for Computational Linguistics, 2018, pp. 57–66.

[153] S. Lappin, H.J. Leass, An algorithm for pronominal anaphora resolution, Comput. Linguist. 20 (4) (1994) 535–561.

[154] J. Bos, Implementing the binding and accommodation theory for anaphora resolution and presupposition projection, Comput. Linguist. 29 (2) (2003) 179–210.

[155] C.-T.J. Huang, On the distribution and reference of empty pronouns, Linguist. Inquiry (1984) 531–574.

[156] G. Chen, Computational Generation of Chinese Noun Phrases (Ph.D. thesis), Utrecht University, 2022.

[157] G. Chen, K. van Deemter, Understanding the use of quantifiers in mandarin, in: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Association for Computational Linguistics, Online only, 2022, pp. 73–80, URL https://aclanthology.org/2022.findings-aacl.7.

[158] G. Chen, F. Same, K. van Deemter, Neural referential form selection: Generalisability and interpretability, Comput. Speech Lang. 79 (2023) 101466.

[159] C. Chen, V. Ng, Chinese zero pronoun resolution: Some recent advances, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 1360–1365.

[160] S. Fligelstone, Developing a scheme for annotating text to show anaphoric relations, in: New Directions in English Language Corpora: Methodology, Results, Software Developments, 1992, pp. 153–170.

[161] R. Passonneau, Instructions for applying discourse reference annotation for multiple applications (DRAMA), 1997, p. 46, Unpublished Manuscript.

[162] L. Hirschman, P. Robinson, J. Burger, M. Vilain, Automating coreference: The role of annotated training data, in: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, 1997, pp. 118–121.

[163] L. Hirschman, N. Chinchor, Appendix F: MUC-7 coreference task definition (version 3.0), in: Seventh Message Understanding Conference, MUC-7, 1998, pp. 1–17.

[164] N.A. Chinchor, B. Sundheim, Message understanding conference (MUC) tests of discourse processing, in: Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, 1995, pp. 21–26.

[165] G.R. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel, The automatic content extraction (ACE) program–tasks, data, and evaluation, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation, 2004, pp. 1–4.

[166] K.v. Deemter, R. Kibble, On coreferring: Coreference in MUC and related annotation schemes, Comput. Linguist. 26 (4) (2000) 629–637.

[167] M. Poesio, F. Bruneseaux, L. Romary, The MATE meta-scheme for coreference in dialogues in multiple languages, in: ACL'99 Workshop Towards Standards and Tools for Discourse Tagging, 1999, pp. 65–74.

[168] M. Poesio, The MATE/GNOME proposals for anaphoric annotation, revisited, in: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue At HLT-NAACL 2004, 2004, pp. 154–162.

[169] I.R. Heim, The Semantics of Definite and Indefinite Noun Phrases, University of Massachusetts Amherst, 1982.

[170] B.L. Webber, A Formal Approach to Discourse Anaphora, Routledge, 2016.

[171] H. Kamp, U. Reyle, From Discourse to Logic: Introduction To Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Vol. 42, Springer Science & Business Media, 2013.

[172] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes, in: Joint Conference on EMNLP and CoNLL-Shared Task, 2012, pp. 1–40.

[173] H.H. Clark, Bridging, in: Theoretical Issues in Natural Language Processing, 1975, pp. 169–174.

[174] B.L. Webber, Discourse deixis: Reference to discourse segments, in: 26th Annual Meeting of the Association for Computational Linguistics, 1988, pp. 113–122.

[175] M. Poesio, R. Artstein, Anaphoric Annotation in the ARRAU Corpus, Technical Report, University of Southern California Los Angeles, 2008.

[176] M. Poesio, Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results, in: Proceedings of the Second International Conference on Language Resources and Evaluation, 2000, pp. 1–8.

[177] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, R. Weischedel, OntoNotes: The 90% solution, in: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 2006, pp. 57–60.

[178] H. Levesque, E. Davis, L. Morgenstern, The Winograd schema challenge, in: Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning, 2012, pp. 552–561.

[179] A. Rahman, V. Ng, Resolving complex cases of definite Pronouns: The winograd schema challenge, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 777–789.

[180] K. Webster, M. Recasens, V. Axelrod, J. Baldridge, Mind the GAP: A balanced corpus of gendered ambiguous pronouns, Trans. Assoc. Comput. Linguist. 6 (2018) 605–617.

[181] L. Hasler, C. Orăsan, K. Naumann, NPs for events: Experiments in coreference annotation, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, 2006, pp. 1167–1172.

[182] A. Cybulska, P. Vossen, Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014, pp. 4545–4552.

[183] M. Poesio, R. Artstein, Anaphoric annotation in the ARRAU corpus, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation, 2008, pp. 1–5.

[184] N.A. Chinchor, Overview of MUC-7, in: Seventh Message Understanding Conference, MUC-7, 1998, pp. 1–4.

[185] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus—a semantically annotated corpus for bio-textmining, Bioinformatics 19 (suppl_1) (2003) i180–i182.

[186] A. Zeldes, The GUM corpus: Creating multilayer resources in the classroom, Language Resources and Evaluation 51 (3) (2017) 581–612.

[187] H. Chen, Z. Fan, H. Lu, A. Yuille, S. Rong, PreCo: A large-scale dataset in preschool vocabulary for coreference resolution, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 172–181.

[188] J. Pearson, R. Stevenson, M. Poesio, The effects of animacy, thematic role, and surface position on the focusing of entities in discourse, in: Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processing, 2001, pp. 1472–1504.

[189] M. Poesio, Associative descriptions and salience: A preliminary investigation, in: Proceedings of the 2003 EACL Workshop on the Computational Treatment of Anaphora, 2003, pp. 31–38.

[190] M. Poesio, M. Alexandrov-Kabadjov, A general-purpose, off the shelf anaphoric resolver, in: Proceedings of Language Resources and Evaluation Conference, 2004, pp. 653–656.

[191] M. Poesio, R. Henschel, J. Hitzeman, R. Kibble, S. Montague, K. van Deemter, Towards an Annotation Scheme for Noun Phrase Generation, Technical Report, European Chapter of the Association for Computational Linguistics, 1999, pp. 1–7.

[192] J. Hitzeman, A.W. Black, P. Taylor, C. Mellish, J. Oberlander, On the use of automatically generated discourse-level information in a concept-to-speech synthesis system, in: 5th International Conference on Spoken Language Processing, 1998, pp. 2763–2766.

[193] K.B. Cohen, A. Lanfranchi, M.J.-y. Choi, M. Bada, W.A. Baumgartner, N. Panteleyeva, K. Verspoor, M. Palmer, L.E. Hunter, Coreference annotation and resolution in the Colorado Richly annotated full text (CRAFT) corpus of biomedical journal articles, BMC Bioinformatics 18 (1) (2017) 1–14.

[194] A. Ghaddar, P. Langlais, WikiCoref: An English coreference-annotated corpus of Wikipedia articles, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016, pp. 136–142.

[195] D. Bamman, O. Lewke, A. Mansoor, An annoated dataset of coreference in English literature, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 44–54.

[196] Y. Hou, K. Markert, M. Strube, Unrestricted bridging resolution, Comput. Linguist. 44 (2) (2018) 237–284.

[197] O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, F. Delogu, K.J. Rodriguez, M. Poesio, Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus, Nat. Lang. Eng. 26 (1) (2020) 95–128.

[198] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1) (1972) 1–191.

[199] E. Davis, L. Morgenstern, C.L. Ortiz, The first Winograd schema challenge at IJCAI-16, AI Mag. 38 (3) (2017) 97–98.

[200] K. Sakaguchi, R.L. Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, Commun. ACM 64 (9) (2021) 99–106.

[201] R. Rudinger, J. Naradowsky, B. Leonard, B. Van Durme, Gender bias in coreference resolution, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 8–14.

[202] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 15–20.

[203] J. Muzerelle, A. Lefeuvre, J.-Y. Antoine, E. Schang, D. Maurel, J. Villaneau, I. Eshkol, ANCOR, the first large french speaking corpus of conversational speech annotated in coreference to be freely available (ANCOR, premier corpus de francais parlé d'envergure annoté en coréférence et distribué librement) [in French], in: Proceedings of TALN 2013 (Volume 2: Short Papers), 2013, pp. 555–563.

[204] M. Taulé, M.A. Martí, M. Recasens, AnCora: Multilevel annotated corpora for Catalan and Spanish, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC'08, 2008, pp. 1–6.

[205] I. Hendrickx, G. Bouma, F. Coppens, W. Daelemans, V. Hoste, G. Kloosterman, A.-M. Mineur, J. Van Der Vloet, J.-L. Verschelde, A coreference corpus and resolution system for dutch, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation, 2008, pp. 1–6.

[206] R. Iida, M. Komachi, K. Inui, Y. Matsumoto, Annotating a Japanese text corpus with predicate-argument and coreference relations, in: Proceedings of the Linguistic Annotation Workshop, 2007, pp. 132–139.

[207] M. Ogrodniczuk, K. Głowińska, M. Kopeć, A. Savary, M. Zawisławska, Polish coreference corpus, in: Language and Technology Conference, Springer, 2013, pp. 215–226.

[208] A. Nedoluzhko, J. Mírovskỳ, E. Fučíková, J. Pergler, Annotation of Coreference in Prague Czech-English Dependency Treebank, Technical report, Technical report 2014/57. Prague: ÚFAL MFF UK, 2014.

[209] H. Telljohann, E. Hinrichs, S. Kübler, R. Kübler, The TüBa-D/Z treebank: Annotating german with a context-free backbone, in: In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Citeseer, 2004, pp. 2229–2232.

[210] S. Martin, The role of salience ranking in anaphora resolution, in: ESSLLI 27 workshop Logic and Probabilistic Methods for Dialog, 2015.

[211] G.A. Miller, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[212] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85.

[213] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The Semantic Web, 2007, pp. 722–735.

[214] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: A collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp. 1247–1250.

[215] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: A large ontology from Wikipedia and WordNet, J. Web Semant. 6 (3) (2008) 203–217.

[216] V. Nastase, M. Strube, B. Boerschinger, C. Zirn, A. Elghafari, WikiNet: A very large scale multi-lingual concept network, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 1015–1022.

[217] P. Singh, The open mind common sense project, KurzweilAI.net 143 (2002) 1–12.

[218] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (5) (2012) 786–791.

[219] A. Emami, A. Trischler, K. Suleman, J.C.K. Cheung, A generalized knowledge hunting framework for the winograd schema challenge, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, 2018, pp. 25–31.

[220] N.S. Moosavi, M. Strube, Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 632–642.

[221] M. Vilain, J.D. Burger, J. Aberdeen, D. Connolly, L. Hirschman, A model-theoretic coreference scoring scheme, in: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995, 1995, pp. 45–52.

[222] A. Bagga, B. Baldwin, Algorithms for scoring coreference chains, in: The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, Vol. 1, 1998, pp. 563–566.

[223] X. Luo, On coreference resolution performance metrics, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 25–32.

[224] S. Kübler, D. Zhekova, Singletons and coreference resolution evaluation, in: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, 2011, pp. 261–267.

[225] X. Luo, S. Pradhan, Evaluation metrics, in: Anaphora Resolution, Springer, 2016, pp. 141–163.

[226] P. Denis, J. Baldridge, Global joint models for coreference resolution and named entity classification, Procesamiento del lenguaje Nat. 42 (2009).

[227] V. Stoyanov, N. Gilbert, C. Cardie, E. Riloff, Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 656–664.

[228] M. Recasens, E. Hovy, BLANC: Implementing the rand index for coreference evaluation, Nat. Lang. Eng. 17 (4) (2011) 485–510.

[229] V. Bartalesi Lenzi, G. Moretti, R. Sprugnoli, CAT: the CELCT annotation tool, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12, 2012, pp. 333–338.

[230] N. Reiter, CorefAnnotator - a new annotation tool for entity references, in: Abstracts of EADH: Data in the Digital Humanities, 2018, pp. 1–4.

[231] B. Oberle, SACR: A Drag-and-Drop Based Tool for Coreference Annotation, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018, pp. 389–394.

[232] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: A web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations At the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.

[233] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, I. Gurevych, The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 2018, pp. 5–9.

[234] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A.V. Tendulkar, F. Leitner, A. Valencia, C. Marcelle, MyMiner: A web application for computer-assisted biocuration and text annotation, Bioinformatics 28 (17) (2012) 2285–2287.

[235] M. Neves, J. Ševa, An extensive review of tools for manual annotation of documents, Brief. Bioinform. 22 (1) (2021) 146–163.

[236] C. Girardi, M. Speranza, R. Sprugnoli, S. Tonelli, Cromer: A tool for cross-document event and entity coreference, in: Ninth International Conference on Language Resources and Evaluation, 2014, pp. 3204–3208.

[237] A. Bornstein, A. Cattan, I. Dagan, CoRefi: A crowd sourcing suite for coreference annotation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 205–215.

[238] J.R. Hobbs, Resolving pronoun references, Lingua 44 (4) (1978) 311–338.

[239] C.L. Sidner, Towards a computational theory of definite anaphora comprehension in english discourse, Technical Report, Massachusetts Inst of Tech Cambridge Artificial Intelligence lab, 1979.

[240] B.J. Grosz, The Representation and Use of Focus in Dialogue Understanding, University of California, Berkeley, 1977.

[241] D. Carter, Interpreting Anaphors in Natural Language Texts, Halsted Press, 1987.

[242] J.R. Hobbs, M. Stickel, P. Martin, D. Edwards, Interpretation as abduction, in: 26th Annual Meeting of the Association for Computational Linguistics, 1988, pp. 95–103.

[243] S. Lappin, H.J. Leass, An algorithm for pronominal anaphora resolution, Comput. Linguist. 20 (4) (1994) 535–561.

[244] S.E. Brennan, M.W. Friedman, C.J. Pollard, A centering approach to Pronouns, in: 25th Annual Meeting of the Association for Computational Linguistics, 1987, pp. 155–162.

[245] D.I. Beaver, The optimization of discourse anaphora, Linguist. Philos. 27 (2004) 3–56.

[246] J.R. Tetreault, A corpus-based evaluation of centering and pronoun resolution, Comput. Linguist. 27 (4) (2001) 507–520.

[247] N. Ge, J. Hale, E. Charniak, A statistical approach to anaphora resolution, in: Sixth Workshop on Very Large Corpora, 1998, pp. 161–170.

[248] B. Baldwin, CogNIAC: High precision coreference with limited knowledge and linguistic resources, in: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, 1997, pp. 38–45.

[249] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, D. Jurafsky, Deterministic coreference resolution based on entity-centric, precision-ranked rules, Comput. Linguist. 39 (4) (2013) 885–916.

[250] M. Kameyama, Recognizing referential links: An information extraction prespective, in: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, 1997, pp. 46–53.

[251] S.M. Harabagiu, S.J. Maiorano, Knowledge-lean coreference resolution and its relation to textual cohesion and coherence, in: The Relation of Discourse/Dialogue Structure and Reference, 1999, pp. 29–38.

[252] T. Liang, D.-S. Wu, Automatic pronominal anaphora resolution in English texts, in: Proceedings of Research on Computational Linguistics Conference, 2003, pp. 111–127.

[253] A. Haghighi, D. Klein, Simple coreference resolution with rich syntactic and semantic features, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 1152–1161.

[254] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, C. Manning, A multi-pass sieve for coreference resolution, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 492–501.

[255] C. Aone, S.W. Bennett, Automated acquisition of anaphora resolution strategies, AAAI 1–7.

[256] W.M. Soon, H.T. Ng, D.C.Y. Lim, A machine learning approach to coreference resolution of noun phrases, Comput. Linguist. 27 (4) (2001) 521–544.

[257] K.W. Church, A stochastic parts program and noun phrase parser for unrestricted text, in: International Conference on Acoustics, Speech, and Signal Processing, 1989, pp. 695–698.

[258] K. Lata, P. Singh, K. Dutta, Mention detection in coreference resolution: Survey, Appl. Intell. (2022) 1–45.

[259] R. Vieira, M. Poesio, An empirically-based system for processing definite descriptions, Comput. Linguist. 26 (4) (2000) 539–593.

[260] S.P. Ponzetto, M. Strube, Exploiting semantic role labeling, WordNet and wikipedia for coreference resolution, in: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006, pp. 192–199.

[261] E. Bengtson, D. Roth, Understanding the value of features for coreference resolution, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 294–303.

[262] V. Ng, C. Cardie, Improving machine learning approaches to coreference resolution, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 104–111.

[263] S.M. Harabagiu, R.C. Bunescu, S.J. Maiorano, Text and knowledge mining for coreference resolution, in: Second Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 1–8.

[264] V. Ng, C. Cardie, Combining sample selection and error-driven pruning for machine learning of coreference rules, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, 2002, pp. 55–62.

[265] M. Strube, S. Rapp, C. Müller, The influence of minimum edit distance on reference resolution, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, 2002, pp. 312–319.

[266] X. Yang, G. Zhou, J. Su, C.L. Tan, Coreference resolution using competition learning approach, in: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 176–183.

[267] J.F. McCarthy, W.G. Lehnert, Using decision trees for conference resolution, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence-Volume 2, 1995, pp. 1050–1055.

[268] H. Lee, M. Surdeanu, D. Jurafsky, A scaffolding approach to coreference resolution integrating statistical and rule-based models, Nat. Lang. Eng. 23 (5) (2017) 733–762.

[269] A.L. Berger, S.A. Della Pietra, V.J. Della Pietra, A maximum entropy approach to natural language processing, Comput. Linguist. 22 (1) (1996) 39–71.

[270] W. Daelemans, J. Zavrel, K. Van Der Sloot, A. Van den Bosch, TiMBL: Tilburg Memory-Based Learner, Tilburg University, 2004.

[271] A. McCallum, B. Wellner, Conditional models of identity uncertainty with application to noun coreference, Adv. Neural Inform. Process. Syst. 17 (2004) 1–8.

[272] A. McCallum, B. Wellner, Object consolidation by graph partitioning with a conditionally-trained distance metric, in: KDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003, pp. 1–6.

[273] C. Nicolae, G. Nicolae, BESTCUT: A graph algorithm for coreference resolution, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 275–283.

[274] V. Ng, Supervised noun phrase coreference research: The first fifteen years, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1396–1411.

[275] C. Cardie, K. Wagstaff, Noun phrase coreference as clustering, in: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999, pp. 82–89.

[276] P. Denis, J. Baldridge, Specialized models and ranking for coreference resolution, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 660–669.

[277] A. Rahman, V. Ng, Narrowing the modeling gap: A cluster-ranking approach to coreference resolution, J. Artificial Intelligence Res. 40 (2011) 469–521.

[278] S. Wiseman, A.M. Rush, S. Shieber, J. Weston, Learning anaphoricity and antecedent ranking features for coreference resolution, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1416–1426.

[279] S. Wiseman, A.M. Rush, S.M. Shieber, Learning global features for coreference resolution, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 994–1004.

[280] K. Clark, C.D. Manning, Improving coreference resolution by learning entity-level distributed representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 643–653.

[281] K. Clark, C.D. Manning, Deep reinforcement learning for mention-ranking coreference models, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2256–2262.

[282] K. Clark, C.D. Manning, Entity-centric coreference resolution with model stacking, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1405–1415.

[283] R. Liu, G. Chen, R. Mao, E. Cambria, A multi-task learning model for gold-two-mention co-reference resolution, in: 2023 International Joint Conference on Neural Networks, 2023, pp. 1–8.

[284] V. Kocijan, O.-M. Camburu, A.-M. Cretu, Y. Yordanov, P. Blunsom, T. Lukasiewicz, WikiCREM: A large unsupervised corpus for coreference resolution, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 4303–4312.

[285] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end neural coreference resolution, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 188–197.

[286] D. Bahdanau, K.H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, 2015, pp. 1–15.

[287] K. Lee, L. He, L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 687–692.

[288] H. Luo, J. Glass, Learning word representations with cross-sentence dependency for end-to-end co-reference resolution, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4829–4833.

[289] R. Zhang, C. Nogueira dos Santos, M. Yasunaga, B. Xiang, D. Radev, Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 102–107.

[290] Y. Kirstain, O. Ram, O. Levy, Coreference resolution without span representations, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 14–19.

[291] W. Wu, F. Wang, A. Yuan, F. Wu, J. Li, CorefQA: Coreference resolution as query-based span prediction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6953–6963.

[292] R. Aralikatte, H. Lent, A.V. Gonzalez, D. Herschcovich, C. Qiu, A. Sandholm, M. Ringaard, A. Søgaard, Rewarding coreference resolvers for being consistent with world knowledge, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 1229–1235.

[293] H. Zhang, Y. Song, Y. Song, D. Yu, Knowledge-aware pronoun coreference resolution, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 867–876.

[294] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases? in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 2463–2473.

[295] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77.

[296] D. Ye, Y. Lin, J. Du, Z. Liu, P. Li, M. Sun, Z. Liu, Coreferential reasoning learning for language representation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7170–7186.

[297] S. Attree, Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 2019, pp. 134–146.

[298] C.L. Yang, P.C. Gordon, R. Hendrick, J.T. Wu, Comprehension of referring expressions in Chinese, Language and Cognitive Processes 14 (5–6) (1999) 715–743.

[299] M. Kameyama, Zero Anaphora: The Case of Japanese (Discourse Aanalysis, Pronouns, Sytax, Computational Llinguistics, Typology), Stanford University, 1985.

[300] M. Okumura, K. Tamura, Zero Pronoun resolution in Japanese discourse based on centering theory, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996, pp. 871–876.

[301] R. Iida, K. Inui, Y. Matsumoto, Zero-anaphora resolution by learning rich syntactic pattern features, ACM Trans. Asian Lang. Inform. Process. (TALIP) 6 (4) (2007) 1–22.

[302] H. Isozaki, T. Hirao, Japanese zero pronoun resolution based on ranking rules and machine learning, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 184–191.

[303] H. Nakaiwa, S. Shirai, S. Ikehara, T. Kawaoka, Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints, in: Proceedings of the AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation, 1995, pp. 99–105.

[304] H. Nakaiwa, S. Shirai, Anaphora resolution of Japanese zero pronouns with deictic reference, in: The 16th International Conference on Computational Linguistics, 1996, pp. 812–817.

[305] K. Seki, A. Fujii, T. Ishikawa, A probabilistic model for Japanese zero pronoun resolution integrating syntactic and semantic features, in: NLPRS, 2001, pp. 403–410.

[306] K. Seki, A. Fujii, T. Ishikawa, A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution, in: Proceedings of the 19th International Conference on Computational Linguistics, 2002, pp. 1–7.

[307] R. Sasano, D. Kawahara, S. Kurohashi, A fully-lexicalized probabilistic model for Japanese zero Anaphora resolution, in: Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 769–776.

[308] R. Sasano, S. Kurohashi, A discriminative approach to Japanese zero Anaphora resolution with large-scale lexicalized case frames, in: Proceedings of 5th International Joint Conference on Natural Language Processing, 2011, pp. 758–766.

[309] D.K. Byron, W. Gegg-Harrison, S.-H. Lee, Resolving zero anaphors and pronouns in Korean, Traitement Autom. des Langues 46 (1) (2006) 91–114.

[310] N.-R. Han, Korean Zero Pronouns: Analysis and Resolution, University of Pennsylvania, 2006.

[311] S. Zhao, H.T. Ng, Identification and resolution of Chinese zero pronouns: A machine learning approach, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 541–550.

[312] F. Kong, G. Zhou, A tree kernel-based unified framework for Chinese zero anaphora resolution, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 882–891.

[313] C. Chen, V. Ng, Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 763–774.

[314] C. Chen, V. Ng, Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 320–326.

[315] A. Aloraini, M. Poesio, Cross-lingual zero pronoun resolution, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 90–98.

[316] C. Chen, V. Ng, Chinese zero pronoun resolution with deep neural networks, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 778–788.

[317] Q. Yin, W. Zhang, Y. Zhang, T. Liu, A deep neural network for Chinese zero pronoun resolution, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3322–3328.

[318] Q. Yin, W. Zhang, Y. Zhang, T. Liu, Chinese zero pronoun resolution: A collaborative filtering-based approach, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19 (1) (2019) 1–20.

[319] Q. Yin, Y. Zhang, W. Zhang, T. Liu, W.Y. Wang, Zero pronoun resolution with attention-based neural network, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 13–23.

[320] Q. Yin, Y. Zhang, W. Zhang, T. Liu, Chinese zero pronoun resolution with deep memory network, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1309–1318.

[321] Q. Yin, Y. Zhang, W.-N. Zhang, T. Liu, W.Y. Wang, Deep reinforcement learning for Chinese zero pronoun resolution, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 569–578.

[322] L. Song, K. Xu, Y. Zhang, J. Chen, D. Yu, ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5429–5434.

[323] S. Chen, B. Gu, J. Qu, Z. Li, A. Liu, L. Zhao, Z. Chen, Tackling zero Pronoun resolution and non-zero coreference resolution jointly, in: Proceedings of the 25th Conference on Computational Natural Language Learning, 2021, pp. 518–527.

[324] A. Aloraini, S. Pradhan, M. Poesio, Joint coreference resolution for zeros and non-zeros in Arabic, 2022, arXiv preprint arXiv:2210.12169.

[325] R. Iida, M. Poesio, A cross-lingual ILP solution to zero anaphora resolution, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 804–813.

[326] T. Liu, Y. Cui, Q. Yin, W.-N. Zhang, S. Wang, G. Hu, Generating and exploiting large-scale pseudo training data for zero pronoun resolution, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 102–111.

[327] A. Aloraini, M. Poesio, Data augmentation methods for anaphoric zero pronouns, in: Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference, 2021, pp. 82–93.

[328] D. Stojanovski, A. Fraser, Coreference and coherence in neural machine translation: A study using oracle experiments, in: Proceedings of the Third Conference on Machine Translation: Research Papers, 2018, pp. 49–60.

[329] D. Saunders, R. Sallis, B. Byrne, Neural machine translation doesn't translate gender coreference right unless you make it, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, 2020, pp. 35–43.
[330] R. Le Nagard, P. Koehn, Aiding pronoun translation with co-reference resolution, in: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics, MATR, 2010, pp. 252–261.
[331] C. Hardmeier, M. Federico, Modelling pronominal anaphora in statistical machine translation, in: IWSLT (International Workshop on Spoken Language Translation), Paris, France; December 2nd and 3rd, 2010, 2010, pp. 283–289.
[332] L. Guillou, Improving Pronoun translation for statistical machine translation, in: Proceedings of the Student Research Workshop At the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 1–10.
[333] L.M. Miculicich, A. Popescu-Belis, Using coreference links to improve spanish-to-english machine translation, in: Proceedings of the 2nd Workshop on Coreference Resolution beyond OntoNOtes, CORBON 2017, 2017, pp. 30–40.
[334] H. Nakaiwa, S. Ikehara, Zero pronoun resolution in a machine translation system by using Japanese to English verbal semantic attributes, in: Third Conference on Applied Natural Language Processing, 1992, pp. 201–208.
[335] X. Tan, S. Kuang, D. Xiong, Detecting and translating dropped pronouns in neural machine translation, in: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8, 2019, pp. 343–354.
[336] L. Wang, Z. Tu, X. Wang, S. Shi, One model to learn both: Zero pronoun prediction and translation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 921–930.
[337] J. Steinberger, M. Poesio, M.A. Kabadjov, K. Ježek, Two uses of Anaphora resolution in summarization, Inf. Process. Manage. 43 (6) (2007) 1663–1680.
[338] S. Bergler, R. Witte, M. Khalife, Z. Li, F. Rudzicz, Using knowledge-poor coreference resolution for text summarization, in: Workshop on Text Summarization, 2003, pp. 1–8.
[339] R. Witte, S. Bergler, Fuzzy coreference resolution for summarization, in: Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization, ARQAS, Citeseer, 2003, pp. 43–50.
[340] S. Sonawane, P. Kulkarni, The role of coreference resolution in extractive summarization, in: 2016 International Conference on Computing, Analytics and Security Trends, CAST, 2016, pp. 351–356.
[341] Z. Liu, K. Shi, N. Chen, Coreference-aware dialogue summarization, in: Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2021, pp. 509–519.
[342] C. Orasan, The influence of pronominal anaphora resolution on term-based summarisation, Recent Adv. Nat. Lang. Process. V: Selected Pap. RANLP (2007) 291–300.
[343] R. Mitkov, R. Evans, C. Orăsan, L.A. Ha, V. Pekar, Anaphora resolution: To what extent does it help NLP applications? in: Anaphora: Analysis, Algorithms and Applications: 6th Discourse Anaphora and Anaphor Resolution Colloquium, 2007, pp. 179–190.
[344] S. Mirkin, I. Dagan, S. Padó, Assessing the role of discourse references in entailment inference, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1209–1219.
[345] L. Bentivogli, P. Clark, I. Dagan, D. Giampiccolo, The fifth PASCAL recognizing textual entailment challenge, in: TAC, Citeseer, 2009, pp. 1–18.
[346] R. Adams, G. Nicolae, C. Nicolae, S. Harabagiu, Textual entailment through extended lexical overlap and lexico-semantic matching, in: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 2007, pp. 119–124.
[347] E. Agichtein, W. Askew, Y. Liu, Combining lexical, syntactic, and semantic evidence for textual entailment classification, in: TAC, 2008, pp. 1–6.
[348] R. Bar-Haim, I. Dagan, S. Mirkin, E. Shnarch, I. Szpektor, J. Berant, I. Greental, Efficient semantic deduction and approximate matching over compact parse forests, in: TAC, 2008, pp. 1–10.
[349] N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, M.-C. de Marneffe, D. Ramage, E. Yeh, C.D. Manning, Learning alignments and leveraging natural logic, in: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 2007, pp. 165–170.
[350] N. Nicolov, F. Salvetti, S. Ivanova, Sentiment analysis: Does coreference matter, in: AISB 2008 Convention Communication, Interaction and Social Intelligence, Vol. 1, 2008, p. 37.
[351] N. Jakob, I. Gurevych, Using anaphora resolution to improve opinion target identification in movie reviews, in: Annual Meeting of the Association for Computational Linguistics, 2010, pp. 263–268.
[352] X. Ding, B. Liu, Resolving object and attribute coreference in opinion mining, in: Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 268–276.
[353] T.T. Le, T.H. Vo, D.T. Mai, T.T. Quan, T.T. Phan, Sentiment analysis using anaphoric coreference resolution and ontology inference, in: International Workshop on Multi-Disciplinary Trends in Artificial Intelligence, 2016, pp. 297–303.
[354] H. Chai, M. Strube, Incorporating centering theory into neural coreference resolution, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 2996–3002.
[355] M. Joshi, O. Levy, L. Zettlemoyer, D. Weld, BERT for coreference resolution: Baselines and analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 5803–5808.
[356] A. Nedoluzhko, M. Novák, M. Popel, Z. Žabokrtskỳ, A. Zeldes, D. Zeman, CorefUD 1.0: Coreference meets universal dependencies, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 4859–4872.
[357] F. Same, G. Chen, K. Van Deemter, Non-neural models matter: A re-evaluation of neural referring expression generation systems, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5554–5567, http://dx.doi.org/10.18653/v1/2022.acl-long.380, URL https://aclanthology.org/2022.acl-long.380.
[358] Y. Li, X. Ma, X. Zhou, P. Cheng, K. He, C. Li, Knowledge enhanced lstm for coreference resolution on biomedical texts, Bioinformatics 37 (17) (2021) 2699–2705.
[359] K. He, B. Mao, X. Zhou, Y. Li, T. Gong, C. Li, J. Wu, Knowledge enhanced coreference resolution via gated attention, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 2287–2293.
[360] R. Aralikatte, M. Lamm, D. Hardt, A. Søgaard, Ellipsis resolution as question answering: An evaluation, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 810–817, http://dx.doi.org/10.18653/v1/2021.eacl-main.68, URL https://aclanthology.org/2021.eacl-main.68.
[361] A. Uma, D. Almanea, M. Poesio, Scaling and disagreements: Bias, noise, and ambiguity, Frontiers in Artificial Intelligence 5 (2022) 1–11.
[362] E. van Miltenburg, W.-T. Lu, E. Krahmer, A. Gatt, G. Chen, L. Li, K. van Deemter, Gradations of error severity in automatic image descriptions, in: Proceedings of the 13th International Conference on Natural Language Generation, 2020, pp. 398–411.
[363] S. Martschat, M. Strube, Recall error analysis for coreference resolution, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 2070–2081.
[364] A.N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, J. Artificial Intelligence Res. 72 (2021) 1385–1470.
[365] J.R. Finkel, T. Grenager, C.D. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 363–370.
[366] N. Jinjie, P. Vlad, Y. Tom, Z. Haicang, C. Erik, HiTKG: Towards goal-oriented conversations via multi-hierarchy learning, in: AAAI Conference on Artificial Intelligence, 2022, pp. 11112–11120.
[367] C. Gao, X. Wang, X. He, Y. Li, Graph neural networks for recommender system, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 1623–1625.
[368] K. He, L. Yao, J. Zhang, Y. Li, C. Li, et al., Construction of genealogical knowledge graphs from obituaries: Multitask neural network extraction system, J. Med. Internet Res. 23 (8) (2021) e25670.
[369] K. He, N. Hong, S. Lapalme-Remis, Y. Lan, M. Huang, C. Li, L. Yao, Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups, Epilep. Behav. 94 (2019) 65–71.
[370] C. Li, X. Xu, G. Zhou, K. He, T. Qi, W. Zhang, F. Tian, Q. Zheng, J. Hu, et al., Implementation of national health informatization in China: Survey about the status quo, JMIR Med. Inform. 7 (1) (2019) e12238.
[371] N. Chinchor, L. Hirschman, D.D. Lewis, Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3), Comput. Linguist. 19 (3) (1993) 409–450.
[372] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 160–167.
[373] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.
[374] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, W. Huang, B. Wen, Y. Liu, Global pointer: Novel efficient span-based approach for named entity recognition, 2022, arXiv preprint arXiv:2208.03054.
[375] Y. Huang, K. He, Y. Wang, X. Zhang, T. Gong, R. Mao, C. Li, COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2022, pp. 2515–2527.
[376] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 50–61.

[377] B. Mao, C. Jia, Y. Huang, K. He, J. Wu, T. Gong, C. Li, Uncertainty-guided mutual consistency training for semi-supervised biomedical relation extraction, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 2318–2325.

[378] N. Ringland, X. Dai, B. Hachey, S. Karimi, C. Paris, J.R. Curran, NNE: A dataset for nested named entity recognition in English newswire, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5176–5181.

[379] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. 34 (1) (2020) 50–70.

[380] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2145–2158.

[381] B. Song, F. Li, Y. Liu, X. Zeng, Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison, Brief. Bioinform. 22 (6) (2021) bbab282.

[382] P. Liu, Y. Guo, F. Wang, G. Li, Chinese named entity recognition: The state of the art, Neurocomputing 473 (2022) 37–53.

[383] Z. Nasar, S.W. Jaffry, M.K. Malik, Named entity recognition and relation extraction: State-of-the-art, ACM Comput. Surv. 54 (1) (2021) 1–39.

[384] E.H. Rosch, Natural categories, Cogn. psychol. 4 (3) (1973) 328–350.

[385] E. Rosch, C.B. Mervis, Family resemblances: Studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605.

[386] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cogn. Psychol. 8 (3) (1976) 382–439.

[387] G. Fauconnier, M. Turner, The Way We Think: Conceptual Blending and the Mind's Hidden Complexities, Basic Books, 2008.

[388] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J.M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Comput. Stand. Interfaces 35 (5) (2013) 482–489.

[389] O. Borrega, M. Taulé, M.A. Martı, What do we mean when we speak about named entities, in: Proceedings of Corpus Linguistics, Citeseer, 2007, pp. 1–27.

[390] S.A. Kripke, Naming and necessity, in: Semantics of Natural Language, Springer, 1972, pp. 253–355.

[391] J. LaPorte, Rigid designators for properties, Philos. Stud. 130 (2) (2006) 321–336.

[392] R. Grishman, B.M. Sundheim, Message understanding conference-6: A brief history, in: Proceeding of the 16th International Conference on Computational Linguistics, 1996, pp. 466–471.

[393] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, S. Liu, A multimodal deep learning framework for predicting drug–drug interaction events, Bioinformatics 36 (15) (2020) 4316–4322.

[394] D. Demner-Fushman, K.W. Fung, P. Do, R.D. Boyce, T.R. Goodwin, Overview of the TAC 2018 Drug-Drug Interaction Extraction from Drug Labels Track, TAC November (2019) 1–12.

[395] C. Walker, S. Strassel, J. Medero, K. Maeda, ACE 2005 multilingual training corpus, Linguistic Data Consortium, Philadelphia 57 (2006) 45.

[396] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C.D. Manning, Position-aware attention and supervised data improve slot filling, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 35–45.

[397] E.T.K. Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning At HLT-NAACL 2003, 2003, pp. 142–147.

[398] A. Stubbs, Ö. Uzuner, Annotating longitudinal clinical narratives for de-identification: The 2014 I2B2/UTHealth corpus, J. Biomed. Inform. 58 (2015) S20–S29.

[399] H. Gurulingappa, A.M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, J. Biomed. Inform. 45 (5) (2012) 885–892.

[400] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920.

[401] L. Derczynski, E. Nichols, M. van Erp, N. Limsopatham, Results of the WNUT2017 shared task on novel and emerging entity recognition, in: Proceedings of the 3rd Workshop on Noisy User-Generated Text, 2017, pp. 140–147.

[402] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, et al., Ontonotes release 5.0, Linguistic Data Consortium, Philadelphia, PA 23 (2013).

[403] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, A. Valencia, Information Retrieval and Text Mining Technologies for Chemistry, Chem. Rev. 117 (12) (2017) 7673–7761.

[404] S. Tedeschi, R. Navigli, MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 801–812.

[405] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Vol. 3180, 2022, pp. 1038–1063.

[406] N. Ringland, X. Dai, B. Hachey, S. Karimi, C. Paris, J.R. Curran, NNE: A dataset for nested named entity recognition in English newswire, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5176–5181.

[407] N. Alvaro, Y. Miyao, N. Collier, et al., TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations, JMIR Public Health Surv. 3 (2) (2017) e6396.

[408] T. Shibuya, E. Hovy, Nested named entity recognition via second-best sequence learning and decoding, Trans. Assoc. Comput. Linguist. 8 (2020) 605–620.

[409] K. Donnelly, et al., SNOMED-CT: The advanced terminology and coding system for eHealth, Stud. Health Technol. Inform. 121 (2006) 279.

[410] C.E. Lipscomb, Medical subject headings (MeSH), Bull. Med. Lib. Assoc. 88 (3) (2000) 265.

[411] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 36 (suppl_1) (2007) D13–D21.

[412] O. Bodenreider, The unified medical language system (UMLS): Integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270.

[413] J. Hirsch, G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle, L. Manchikanti, ICD-10: History and context, Am. J. Neuroradiol. 37 (4) (2016) 596–599.

[414] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. data 3 (1) (2016) 1–9.

[415] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–D1082.

[416] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M.I. Stefan, et al., BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models, BMC Syst. Biol. 4 (1) (2010) 1–14.

[417] D. Ahlers, Assessment of the accuracy of GeoNames gazetteer data, in: Proceedings of the 7th Workshop on Geographic Information Retrieval, 2013, pp. 74–81.

[418] S. Branahl, Das EDGAR (Electronic Data Gathering, Analysis and Retrieval) System Der SEC Und Seine Bedeutung FÜR Die Bereitstellung Von Rechnungslegungsinformationen, diplom. de, 1998.

[419] J. Hu, Z. Li, B. Xu, An approach of ontology based knowledge base construction for chinese K12 education, in: 2016 First International Conference on Multimedia and Image Processing, 2016, pp. 83–88.

[420] K. Bontcheva, H. Cunningham, I. Roberts, A. Roberts, V. Tablan, N. Aswani, G. Gorrell, GATE Teamware: A web-based, collaborative text annotation framework, Lang. Res. Eval. 47 (4) (2013) 1007–1029.

[421] K. Rim, MAE2: Portable annotation tool for general natural language use, in: Proc 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, 2016, pp. 75–80.

[422] D. Ferrucci, A. Lally, UIMA: An architectural approach to unstructured information processing in the corporate research environment, Nat. Lang. Eng. 10 (3–4) (2004) 327–348.

[423] A. Katiyar, C. Cardie, Nested named entity recognition revisited, in: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 861–871.

[424] J. Straková, M. Straka, J. Hajic, Neural architectures for nested NER through linearization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5326–5331.

[425] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, J. Li, A unified MRC framework for named entity recognition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5849–5859.

[426] Q. Liu, R. Mao, X. Geng, E. Cambria, Semantic matching in machine reading comprehension: An empirical study, Inf. Process. Manage. 60 (2) (2023) 103145.

[427] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, GPT understands, too, 2021, arXiv preprint arXiv:2103.10385.

[428] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, X. Qiu, A unified generative framework for various NER subtasks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5808–5822.

[429] S. Skylaki, A. Oskooei, O. Bari, N. Herger, Z. Kriegman, Named entity recognition in the legal domain using a pointer generator network, 2020, arXiv preprint arXiv:2012.09936.

[430] H. Fei, D. Ji, B. Li, Y. Liu, Y. Ren, F. Li, Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 14, 2021, pp. 12785–12793.

[431] S. Yang, K. Tu, Bottom-up constituency parsing and nested named entity recognition with pointer networks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2403–2416.

[432] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, no. 01, 2019, pp. 3558–3565.

[433] J.R. Finkel, C.D. Manning, Nested named entity recognition, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 141–150.

[434] A.O. Muis, W. Lu, Labeling gaps between words: Recognizing overlapping mentions with mention separators, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2608–2618.

[435] B. Wang, W. Lu, Neural segmental hypergraphs for overlapping mention recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 204–214.

[436] Q. Wan, L. Wei, X. Chen, J. Liu, A region-based hypergraph network for joint entity-relation extraction, Knowl.-Based Syst. 228 (2021) 107298.

[437] Y. Yan, S. Song, Local hypergraph-based nested named entity recognition as query-based sequence labeling, 2022, arXiv preprint arXiv:2204.11467.

[438] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4080–4090.

[439] A. Fritzler, V. Logacheva, M. Kretov, Few-shot classification in named entity recognition task, in: Proceedings of the ACM Symposium on Applied Computing, 2019, pp. 993–1000.

[440] Y. Yang, A. Katiyar, Simple and effective few-shot named entity recognition with structured nearest neighbor learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 6365–6375.

[441] S. Wiseman, K. Stratos, Label-agnostic sequence labeling by copying nearest neighbors, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5363–5369.

[442] S.S.S. Das, A. Katiyar, R. Passonneau, R. Zhang, CONTaiNER: Few-shot named entity recognition via contrastive learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6338–6353.

[443] K. He, Y. Huang, R. Mao, T. Gong, C. Li, E. Cambria, Virtual prompt pre-training for prototype-based few-shot relation extraction, Expert Syst. Appl. 213 (2023) 118927.

[444] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Trans. Affect. Comput. (2022) 1–11.

[445] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2339–2352.

[446] L. Cui, Y. Wu, J. Liu, S. Yang, Y. Zhang, Template-based named entity recognition using BART, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1835–1845.

[447] R. Ma, X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, X. Huang, Template-free prompt tuning for few-shot NER, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 5721–5732.

[448] Y. Huang, K. He, Y. Wang, X. Zhang, T. Gong, R. Mao, C. Li, COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 2515–2527.

[449] M. Miwa, M. Bansal, End-to-end relation extraction using LSTMs on sequences and tree structures, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1105–1116.

[450] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, H. Wang, ERNIE 2.0: A continual pre-training framework for language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 05, 2020, pp. 8968–8975.

[451] Z. Yan, C. Zhang, J. Fu, Q. Zhang, Z. Wei, A partition filter network for joint entity and relation extraction, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 185–197.

[452] M. Miwa, Y. Sasaki, Modeling joint entity and relation extraction with table representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1858–1869.

[453] P. Gupta, H. Schütze, B. Andrassy, Table filling multi-task recurrent neural network for joint entity and relation extraction, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2537–2547.

[454] M. Zhang, Y. Zhang, G. Fu, End-to-end neural relation extraction with global optimization, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1730–1740.

[455] F. Ren, L. Zhang, S. Yin, X. Zhao, S. Liu, B. Li, Y. Liu, A novel global feature-oriented relational triple extraction model based on table filling, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 2646–2656.

[456] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, B. Xu, Joint extraction of entities and relations based on a novel tagging scheme, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1227–1236.

[457] B. Yu, Z. Zhang, X. Shu, T. Liu, Y. Wang, B. Wang, S. Li, Joint extraction of entities and relations based on a novel decomposition strategy, in: ECAI 2020, IOS Press, 2020, pp. 2282–2289.

[458] Z. Wei, J. Su, Y. Wang, Y. Tian, Y. Chang, A novel hierarchical binary tagging framework for joint extraction of entities and relations, 2019, arXiv preprint arXiv:1909.03227.

[459] Y. Yao, Z. Zhang, Y. Xu, C. Li, Data augmentation for few-shot knowledge graph completion from hierarchical perspective, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 2494–2503.

[460] K. He, J. Wu, X. Ma, C. Zhang, M. Huang, C. Li, L. Yao, Extracting kinship from obituary to enhance electronic health records for genetic research, in: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, 2019, pp. 1–10.

[461] T. Jiang, Q. Zeng, T. Zhao, B. Qin, T. Liu, N.V. Chawla, M. Jiang, Biomedical knowledge graphs construction from conditional statements, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (3) (2020) 823–835.

[462] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang, et al., Real-world data medical knowledge graph: Construction and applications, Artif. Intell. Med. 103 (2020) 101817.

[463] S. Silvestri, F. Gargiulo, M. Ciampi, Iterative annotation of biomedical ner corpora with deep neural networks and knowledge bases, Appl. Sci. 12 (12) (2022) 5775.

[464] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, BioNLP 2019 (2019) 58.

[465] S. Shafqat, H. Majeed, Q. Javaid, H.F. Ahmad, Standard NER tagging scheme for big data healthcare analytics built on unified medical corpora, J. Artif. Intell. Technol. 2 (4) (2022) 152–157.

[466] S. Tedeschi, S. Conia, F. Cecconi, R. Navigli, Named entity recognition for entity linking: What works and what's next, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2584–2596.

[467] Z. Batmaz, A. Yurekli, A. Bilge, C. Kaleli, A review on deep learning for recommender systems: Challenges and remedies, Artif. Intell. Rev. 52 (1) (2019) 1–37.

[468] J.-D. Kim, J. Son, D.-K. Baik, CA 5W1H onto: Ontological context-aware model based on 5W1H, Int. J. Distrib. Sens. Netw. 8 (3) (2012) 247346.

[469] K. Zhou, W.X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1006–1014.

[470] J. Wu, K. He, R. Mao, C. Li, E. Cambria, Megacare: Knowledge-guided multi-view hypergraph predictive framework for healthcare, Inf. Fusion 100 (2023) 101939.

[471] A. Iovine, F. Narducci, G. Semeraro, Conversational recommender systems and natural language:: A study through the ConveRSE framework, Decis. Support Syst. 131 (2020) 113250.

[472] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, Exploring high-order user preference on the knowledge graph for recommender systems, ACM Trans. Inf. Syst. 37 (3) (2019) 1–26.

[473] C. Upadhyay, H. Abu-Rasheed, C. Weber, M. Fathi, Explainable job-posting recommendations using knowledge graphs and named entity recognition, in: 2021 IEEE International Conference on Systems, Man, and Cybernetics, 2021, pp. 3291–3296.

[474] J. Ni, T. Young, V. Pandelea, F. Xue, E. Cambria, Recent advances in deep learning based dialogue systems: A systematic survey, Artif. Intell. Rev. (2022) 3055–3155.

[475] X. Li, Y.-N. Chen, L. Li, J. Gao, A. Celikyilmaz, End-to-end task-completion neural dialogue systems, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 733–743.

[476] W.A. Abro, A. Aicher, N. Rach, S. Ultes, W. Minker, G. Qi, Natural language understanding for argumentative dialogue systems in the opinion building domain, Knowl.-Based Syst. 242 (2022) 108318.

[477] E. Dimitrakis, K. Sgontzos, Y. Tzitzikas, A survey on question answering systems over linked data and documents, J. Intell. Inf. Syst. 55 (2) (2020) 233–259.

[478] J. Zhang, Y. Yang, C. Chen, L. He, Z. Yu, KERS: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1092–1101.

[479] M. Kaya, H.c. Bilge, Deep metric learning: A survey, Symmetry 11 (9) (2019) 1066.

[480] B. Kulis, et al., Metric learning: A survey, Found. Trends® Mach. Learn. 5 (4) (2013) 287–364.

[481] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 61–68.

[482] F. Li, M. Zhang, G. Fu, D. Ji, A neural joint model for entity and relation extraction from biomedical text, BMC Bioinformatics 18 (1) (2017) 1–11.

[483] I. Bekoulis, J. Deleu, T. Demeester, C. Develder, Joint entity recognition and relation extraction as a multi-head selection problem, Expert Syst. Appl. 114 (2018) 34–45.

[484] Y. Ma, T. Hiraoka, N. Okazaki, Named entity recognition and relation extraction using enhanced table filling by contextualized representations, J. Nat. Lang. Process. 29 (1) (2022) 187–223.

[485] G. Adomavicius, A. Tuzhilin, Context-aware recommender systems, in: Recommender Systems Handbook, 2011, pp. 217–253.

[486] C.-S. Wu, S.C. Hoi, R. Socher, C. Xiong, TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 917–929.

[487] Y. Wang, Y. Guo, S. Zhu, Slot attention with value normalization for multi-domain dialogue state tracking, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 3019–3028.

[488] A. Kim, H.-J. Song, S.-B. Park, et al., A two-step neural dialog state tracker for task-oriented dialog processing, Comput. Intell. Neurosci. 2018 (2018) 1–12.

[489] P. Hohenecker, F. Mtumbuka, V. Kocijan, T. Lukasiewicz, Systematic comparison of neural architectures and training approaches for open information extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 8554–8565.

[490] K. Kolluru, V. Adlakha, S. Aggarwal, S. Chakrabarti, et al., OpenIE6: Iterative grid labeling and coordination analysis for open information extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 3748–3761.

[491] A. Rahimi, Y. Li, T. Cohn, Massively multilingual transfer for NER, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 151–164.

[492] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, R. Navigli, WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER, in: Findings of the Association for Computational Linguistics: EMNLP, 2021, pp. 2521–2533.

[493] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A continual learning survey: Defying forgetting in classification tasks, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3366–3385.

[494] N. Monaikul, G. Castellucci, S. Filice, O. Rokhlenko, Continual learning for named entity recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 15, 2021, pp. 13570–13577.

[495] Y. Xia, Q. Wang, Y. Lyu, Y. Zhu, W. Wu, S. Li, D. Dai, Learn and review: Enhancing continual named entity recognition via reviewing synthetic samples, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 2291–2300.

[496] S. Vijay, A. Priyanshu, NERDA-Con: Extending NER models for continual learning–Integrating distinct tasks and updating distribution shifts, 2022, arXiv preprint arXiv:2206.14607.

[497] K. He, R. Mao, T. Gong, E. Cambria, C. Li, JCBIE: A joint continual learning neural network for biomedical information extraction, BMC Bioinformatics 23 (1) (2022) 1–20.

[498] C.A. Montgomery, Concept extraction, Am. J. Comput. Linguist. 8 (2) (1982) 70–73.

[499] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, D. Delen, Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, Academic Press, 2012.

[500] Z. Alami Merrouni, B. Frikh, B. Ouhbi, Automatic keyphrase extraction: A survey and trends, J. Intell. Inf. Syst. 54 (2020) 391–424.

[501] C. Havasi, R. Speer, ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge, in: Recent Advances in Natural Language Processing, 2007, pp. 27–29.

[502] R. Snow, D. Jurafsky, A.Y. Ng, Semantic taxonomy induction from heterogenous evidence, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006, pp. 801–808.

[503] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3829–3839.

[504] L. Pan, X. Wang, C. Li, J. Li, J. Tang, Course concept extraction in MOOCs via embedding-based graph propagation, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 875–884.

[505] Y. Liu, H. Wu, Z. Huang, H. Wang, J. Ma, Q. Liu, E. Chen, H. Tao, K. Rui, Technical phrase extraction for patent mining: A multi-level approach, in: 20th IEEE International Conference on Data Mining, 2020, pp. 1142–1147.

[506] C. Kartik Detroja, B.S.B. Bhensdadia, A survey on relation extraction, Intell. Syst. Appl. (2023).

[507] C. Xiong, R. Power, J. Callan, Explicit semantic ranking for academic search via knowledge graph embedding, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 1271–1279.

[508] Y. Fang, Y. Zhang, Data-efficient concept extraction from pre-trained language models for commonsense explanation generation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics, EMNLP, 2022, pp. 5883–5893.

[509] D.L. Medin, M.M. Schaffer, Context theory of classification learning, Psychol. Rev. 85 (3) (1978) 207.

[510] A. Wierzbicka, Semantic Primitives, Athena um-Verl, 1972.

[511] P. Gardenfors, Conceptual Spaces: The Geometry of Thought, MIT Press, 2004.

[512] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003, pp. 1–8.

[513] T.D. Nguyen, M. Kan, Keyphrase extraction in scientific publications, in: D.H. Goh, T.H. Cao, I. Sølvberg, E.M. Rasmussen (Eds.), 10th International Conference on Asian Digital Libraries, in: Lecture Notes in Computer Science, vol.4822, 2007, pp. 317–326.

[514] M. Krapivin, A. Autaeu, M. Marchese, Large Dataset for Keyphrases Extraction, Technical Report, University of Trento, 2009, pp. 1–4.

[515] S.N. Kim, O. Medelyan, M. Kan, T. Baldwin, SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles, in: K. Erk, C. Strapparava (Eds.), Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 21–26.

[516] Q. Zhang, Y. Wang, Y. Gong, X. Huang, Keyphrase extraction using deep recurrent neural networks on Twitter, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 836–845.

[517] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, Y. Chi, Deep keyphrase generation, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 582–592.

[518] Y. Wang, Q. Liu, C. Qin, T. Xu, Y. Wang, E. Chen, H. Xiong, Exploiting topic-based adversarial neural network for cross-domain keyphrase extraction, in: IEEE International Conference on Data Mining, 2018, pp. 597–606.

[519] K. Li, H. Zha, Y. Su, X. Yan, Concept mining via embedding, in: IEEE International Conference on Data Mining, 2018, pp. 267–276.

[520] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, M. Verhagen, SemEval-2016 Task 12: Clinical TempEval, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval-2016, 2016, pp. 1052–1062.

[521] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 I2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.

[522] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, 2018 N2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc. 27 (1) (2020) 3–12.

[523] S. Gehrmann, F. Dernoncourt, Y. Li, E.T. Carlson, J.T. Wu, J. Welt, J. Foote Jr., E.T. Moseley, D.W. Grant, P.D. Tyler, et al., Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, PLoS One 13 (2) (2018) e0192360.

[524] L. Pan, X. Wang, C. Li, J. Li, J. Tang, Course concept extraction in MOOCs via embedding-based graph propagation, in: G. Kondrak, T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017, pp. 875–884.

[525] X. Huang, Q. Liu, C. Wang, H. Han, J. Ma, E. Chen, Y. Su, S. Wang, Constructing educational concept maps with multiple relationships from multi-source data, in: J. Wang, K. Shim, X. Wu (Eds.), 2019 IEEE International Conference on Data Mining, IEEE, 2019, pp. 1108–1113.

[526] J. Chen, X. Zhang, Y. Wu, Z. Yan, Z. Li, Keyphrase generation with correlation constraints, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4057–4066.

[527] R.A. Al-Zaidy, C. Caragea, C.L. Giles, Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents, in: L. Liu, R.W. White, A. Mantrach, F. Silvestri, J.J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, ACM, 2019, pp. 2551–2557.

[528] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: C.E. Brodley, A.P. Danyluk (Eds.), Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 282–289.

[529] S. Fang, Z. Huang, M. He, S. Tong, X. Huang, Y. Liu, J. Huang, Q. Liu, Guided attention network for concept extraction, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 1449–1455.

[530] J. Gu, Z. Lu, H. Li, V.O.K. Li, Incorporating copying mechanism in sequence-to-sequence learning, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1631—1640.

[531] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, Modeling coverage for neural machine translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 76—85.

[532] H. Ye, L. Wang, Semi-supervised learning for neural keyphrase generation, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4142–4153.

[533] W. Chen, Y. Gao, J. Zhang, I. King, M.R. Lyu, Title-guided encoding for keyphrase generation, in: The Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 6268–6275.

[534] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, no. 1, 2017, pp. 4444–4451.

[535] E. Cambria, D. Olsher, D. Rajagopal, SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, no. 1, 2014, pp. 1515–1521.

[536] E. Cambria, S. Poria, R. Bajpai, B. Schuller, SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives, in: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2666–2677.

[537] E. Cambria, R. Mao, S. Han, Q. Liu, Sentic parser: A graph-based approach to concept extraction for sentiment analysis, in: IEEE International Conference on Data Mining Workshops, 2022, pp. 413–420.

[538] M. Ge, R. Mao, E. Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, no. 10, 2022, pp. 10681–10689.

[539] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, 1980.

[540] R. Mao, X. Li, K. He, M. Ge, E. Cambria, MetaPro Online: A computational metaphor processing online system, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2023, pp. 127–135.

[541] P. Li, H. Huang, UTA DLNLP at SemEval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports, in: S. Bethard, D.M. Cer, M. Carpuat, D. Jurgens, P. Nakov, T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation, 2016, pp. 1268–1273.

[542] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (2) (2017) 53–61.

[543] X. Yang, J. Bian, W.R. Hogan, Y. Wu, Clinical concept extraction using transformers, J. Am. Medical Informatics Assoc. 27 (12) (2020) 1935–1942.

[544] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020, pp. 1–17.

[545] K. Clark, M. Luong, Q.V. Le, C.D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: 8th International Conference on Learning Representations, 2020, pp. 1–18.

[546] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015, pp. 1–10, CoRR abs/1508.01991.

[547] L. Lange, H. Adel, J. Strötgen, Closing the gap: Joint de-identification and concept extraction in the clinical domain, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 6945–6952.

[548] X. Wang, W. Feng, J. Tang, Q. Zhong, Course concept extraction in MOOC via explicit/implicit representation, in: Third IEEE International Conference on Data Science in Cyberspace, 2018, pp. 339–345.

[549] W. Li, L. Zhu, R. Mao, E. Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 13121–13129.

[550] S. Han, R. Mao, E. Cambria, Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 94–104.

[551] Q. Liu, Z. Huang, Z. Huang, C. Liu, E. Chen, Y. Su, G. Hu, Finding similar exercises in online education systems, in: Y. Guo, F. Farooq (Eds.), Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1821–1830.

[552] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, (1) 2018, pp. 4970–4977.

[553] L. Huang, Z. Ye, J. Qin, L. Lin, X. Liang, GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 9230–9240.

[554] C.-W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2122–2132.

[555] N.M. Seel, Encyclopedia of the Sciences of Learning, Springer Science & Business Media, 2011.

[556] S. Xiong, Y. Tan, G. Wang, Explore visual concept formation for image classification, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, 139, PMLR, 2021, pp. 11470–11479.

[557] A. Paivio, Abstractness, imagery, and meaningfulness in paired-associate learning, J. Verbal Learn. Verbal Behav. 4 (1) (1965) 32–38.

[558] G. Löhr, What are abstract concepts? On lexical ambiguity and concreteness ratings, Rev. Philos. Psychol. 13 (3) (2022) 549–566.

[559] J. Wiebe, Tracking point of view in narrative, Comput. Linguist. 20 (2) (1994) 233–287.

[560] B. Liu, et al., Sentiment analysis and subjectivity, Handb. Nat. Lang. Process. 2 (2010) (2010) 627–666.

[561] E.C. Traugott, Revisiting subjectification and intersubjectification, Subjectif. Intersubjectif. Grammatical. 29 (2010) 71.

[562] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112.

[563] S.-M. Kim, E. Hovy, Automatic detection of opinion bearing words and sentences, in: Companion Volume To the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts, 2005, pp. 61–66.

[564] K. He, R. Mao, T. Gong, C. Li, E. Cambria, Meta-based self-training and re-weighting for aspect-based sentiment analysis, IEEE Trans. Affect. Comput. (2022).

[565] H. Bao, K. He, X. Yin, X. Li, X. Bao, H. Zhang, J. Wu, Z. Gao, Bert-based meta-learning approach with looking back for sentiment analysis of literary book reviews, in: Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10, Springer, 2021, pp. 235–247.

[566] T. Wilson, J. Wiebe, R. Hwa, Just how mad are you? Finding strong and weak opinion clauses, in: AAAI, 4, 2004, pp. 761–769.

[567] H.N.T. Xuan, A.C. Le, et al., Linguistic features for subjectivity classification, in: 2012 International Conference on Asian Language Processing, IEEE, 2012, pp. 17–20.

[568] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 271–278.

[569] P. Biyani, S. Bhatia, C. Caragea, P. Mitra, Using non-lexical features for identifying factual and opinionative threads in online forums, Knowl.-Based Syst. 69 (2014) 170–178.

[570] S. Rustamov, E. Mustafayev, M.A. Clements, Sentence-level subjectivity detection using neuro-fuzzy models, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 108–114.

[571] A.-z. Kamil, S. Rustamov, M.A. Clements, E. Mustafayev, Adaptive neuro-fuzzy inference system for classification of texts, in: Recent Developments and the New Direction in Soft-Computing Foundations and Applications, Springer, 2018, pp. 63–70.

[572] W. Zhang, C. Yu, W. Meng, Opinion retrieval from blogs, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, 2007, pp. 831–840.

[573] J. Wiebe, E. Riloff, Finding mutual benefit between subjectivity analysis and information extraction, IEEE Trans. Affect. Comput. 2 (4) (2011) 175–191.

[574] W. Wang, L. He, Y.J. Wu, M. Goh, Signaling persuasion in crowdfunding entrepreneurial narratives: the subjectivity vs objectivity debate, Comput. Hum. Behav. 114 (2021) 106576.

[575] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, et al., Learning sentiment-specific word embedding for Twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Citeseer, 2014, pp. 1555–1565.

[576] A. Al Hamoud, A. Hoenig, K. Roy, Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models, J. King Saud Univ.-Comput. Inform. Sci. (2022) 7975–7987.

[577] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 129–136.

[578] B. Li, Y. Liu, E. Agichtein, CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 937–946.

[579] A. Banfield, Unspeakable Sentences (Routledge Revivals): Narration and Representation in the Language of Fiction, Routledge, 2014.

[580] J.M. Wiebe, Recognizing Subjective Sentences: A Computational Investigation of Narrative Text, State University of New York at Buffalo, 1990.
[581] J. Wiebe, et al., Learning subjective adjectives from corpora, AAAI/IAAI 20 (2000).
[582] J.L. Austin, How to Do Things with Words, Oxford University Press, 1975.
[583] R. Mao, C. Lin, F. Guerin, End-to-end sequential metaphor identification inspired by linguistic theories, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 3888–3898.
[584] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation 39 (2) (2005) 165–210.
[585] T. Wilson, Annotating subjective content in meetings, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008, pp. 2738–2745.
[586] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, T. Wilson, SemEval-2013 task 2: Sentiment analysis in Twitter, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEVal 2013, 2013, pp. 312–320.
[587] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 976–983.
[588] C. Banea, R. Mihalcea, J. Wiebe, Multilingual subjectivity: Are more languages better? in: Proceedings of the 23rd International Conference on Computational Linguistics, Coling 2010, 2010, pp. 28–36.
[589] S. Somasundaran, J. Wiebe, Recognizing stances in ideological on-line debates, in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 116–124.
[590] L. Barbosa, J. Feng, Robust sentiment detection on Twitter from biased and noisy data, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 36–44.
[591] J. Serrano-Guerrero, J.A. Olivas, F.P. Romero, E. Herrera-Viedma, Sentiment analysis: A review and comparative analysis of web services, Inform. Sci. 311 (2015) 18–38.
[592] A. Khatua, E. Cambria, S.S. Ho, J.C. Na, Deciphering public opinion of nuclear energy on Twitter, in: 2020 International Joint Conference on Neural Networks, 2020, pp. 1–8.
[593] R. Satapathy, I. Chaturvedi, E. Cambria, S.S. Ho, J.C. Na, Subjectivity detection in nuclear energy tweets, Computación y Sistemas 21 (4) (2017) 657–664.
[594] J. Villena, J. García-Morera, M. García-Cumbreras, E. Martínez-Cámara, M. Martín-Valdivia, L. López, Overview of TASS 2015, in: Proceedings of TASS 2015: Workshop on Sentiment Analysis At SEPLN Co-Located with 31st SEPLN Conference, 2015, pp. 13–21.
[595] P. Chesley, B. Vincent, L. Xu, R.K. Srihari, Using verbs and adjectives to automatically classify blog sentiment, Training 580 (263) (2006) 233–241.
[596] G. Murray, G. Carenini, Subjectivity detection in spoken and written conversations, Nat. Lang. Eng. 17 (3) (2011) 397–418.
[597] J. Ulrich, G. Murray, G. Carenini, A publicly available annotated corpus for supervised email summarization, in: AAAI08 Email Workshop, 2008, pp. 1–6.
[598] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al., The AMI meeting corpus, in: Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, Vol. 88, 2005, p. 100.
[599] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.
[600] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, 2011, pp. 169–176.
[601] C. Hube, B. Fetahu, Detecting biased statements in Wikipedia, in: Companion Proceedings of the Web Conference 2018, 2018, pp. 1779–1786.
[602] R. Pryzant, R.D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, D. Yang, Automatically neutralizing subjective bias in text, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 01, 2020, pp. 480–489.
[603] P.J. Stone, D.C. Dunphy, M.S. Smith, The General Inquirer: A Computer Approach to Content Analysis, MIT Press, 1966.
[604] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2005, pp. 347–354.
[605] C. Strapparava, A. Valitutti, WordNet-Affect: An affective extension of WordNet, in: Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004, pp. 1083–1086.
[606] E. Riloff, Automatically generating extraction patterns from untagged text, in: Proceedings of the National Conference on Artificial Intelligence, 1996, pp. 1044–1049.
[607] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2005, pp. 486–497.
[608] F. Benamara, B. Chardon, Y. Mathieu, V. Popescu, Towards context-based subjectivity analysis, in: Proceedings of 5th International Joint Conference on Natural Language Processing, 2011, pp. 1180–1188.
[609] N. Asher, A. Lascarides, Logics of Conversation, Cambridge University Press, 2003.
[610] R. Remus, Improving sentence-level subjectivity classification through readability measurement, in: Proceedings of the 18th Nordic Conference of Computational Linguistics, 2011, pp. 168–174.
[611] E.A. Smith, Devereux readability index, J. Educ. Res. 54 (8) (1961) 298–303.
[612] I.E. Fang, The "easy listening formula", J. Broadcast. Electron. Media 11 (1) (1966) 63–68.
[613] J. Sixto, A. Almeida, D. López-de Ipiña, An approach to subjectivity detection on Twitter using the structured information, in: International Conference on Computational Collective Intelligence, Springer, 2016, pp. 121–130.
[614] J.M. Cotelo, F. Cruz, F.J. Ortega, J.A. Troyano, Explorando Twitter mediante la integración de información estructurada y no estructurada, Procesamiento del Lenguaje Nat. (55) (2015) 75–82.
[615] H. Keshavarz, M. Saniee Abadeh, MHSublex: Using metaheuristic methods for subjectivity classification of microblogs, J. AI Data Mining 6 (2) (2018) 341–353.
[616] C. Akkaya, J. Wiebe, R. Mihalcea, Subjectivity word sense disambiguation, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 190–199.
[617] A. Kilgarriff, M. Palmer, Introduction to the special issue on SENSEVAL, Comput. Humanit. 34 (1) (2000) 1–13.
[618] J. Preiss, D. Yarowsky, Proceedings of SENSEVAL-2 second international workshop on evaluating word sense disambiguation systems, in: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, 2001, pp. 1–163.
[619] K. Litkowski, SensEval-3 task: Word sense disambiguation of wordnet glosses, in: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004, pp. 13–16.
[620] R. Ortega, A. Fonseca, Y. Gutiérrez, A. Montoyo, Improving subjectivity detection using unsupervised subjectivity word sense disambiguation, Procesamiento del Lenguaje Nat. 51 (2013) 179–186.
[621] H. Anaya-Sánchez, A. Pons-Porrata, R. Berlanga-Llavori, Word sense disambiguation based on word sense clustering, in: Advances in Artificial Intelligence-IBERAMIA-SBIA 2006, Springer, 2006, pp. 472–481.
[622] M. Helmi, S.M.T. AlModarresi, Human activity recognition using a fuzzy inference system, in: 2009 IEEE International Conference on Fuzzy Systems, 2009, pp. 1897–1902.
[623] R. Fullér, Neural Fuzzy Systems, Citeseer, 1995.
[624] S. Rustamov, A hybrid system for subjectivity analysis, Adv. Fuzzy Syst. 2018 (2018).
[625] S. Wang, C. Manning, Fast dropout training, in: International Conference on Machine Learning, PMLR, 2013, pp. 118–126.
[626] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, pp. 1–18, arXiv preprint arXiv:1207.0580.
[627] S.I. Wang, C.D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012, pp. 90–94.
[628] Y. He, Bayesian Models for Sentence-Level Subjectivity Detection, Technical Report, Citeseer, Knowledge Media Institute, 2010.
[629] A. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142–150.
[630] C. Lin, Y. He, R. Everson, Sentence subjectivity detection with weakly-supervised learning, in: Proceedings of 5th International Joint Conference on Natural Language Processing, 2011, pp. 1153–1161.
[631] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 655–665.
[632] H. Zhao, Z. Lu, P. Poupart, Self-adaptive hierarchical sentence model, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 4069–4076.
[633] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in: D. Wu, M. Carpuat, X. Carreras, E.M. Vecchi (Eds.), Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014, pp. 103–111.
[634] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Bayesian network based extreme learning machine for subjectivity detection, J. Franklin Inst. B 355 (4) (2018) 1780–1797.
[635] E. Soria-Olivas, J. Gomez-Sanchis, J.D. Martin, J. Vila-Frances, M. Martinez, J.R. Magdalena, A.J. Serrano, BELM: Bayesian extreme learning machine, IEEE Trans. Neural Netw. 22 (3) (2011) 505–509.
[636] B. Frénay, M. Verleysen, Reinforced extreme learning machines for fast robust regression in the presence of outliers, IEEE Trans. Cybern. 46 (12) (2015) 3351–3363.

[637] Y. Kim, Convolutional neural networks for sentence classification, 2014, pp. 1–6, arXiv.

[638] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4487–4496.

[639] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune BERT for text classification? in: China National Conference on Chinese Computational Linguistics, Springer, 2019, pp. 194–206.

[640] R. Mao, X. Li, Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 15, 2021, pp. 13534–13542.

[641] H. Huo, M. Iwaihara, Utilizing BERT pretrained models with various fine-tune methods for subjectivity detection, in: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Springer, 2020, pp. 270–284.

[642] R. Satapathy, S.R. Pardeshi, E. Cambria, Polarity and subjectivity detection with multitask learning and BERT embedding, Future Internet 14 (7) (2022) 191–201.

[643] R. Socher, D. Chen, C.D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, Adv. Neural Inf. Process. Syst. 26 (2013).

[644] M. Crawshaw, Multi-task learning with deep neural networks: A survey, 2020, pp. 1–43, arXiv preprint arXiv:2009.09796.

[645] S. Sagnika, B.S.P. Mishra, S.K. Meher, An attention-based CNN-LSTM model for subjectivity detection in opinion-mining, Neural Comput. Appl. 33 (24) (2021) 17425–17438.

[646] S. Sagnika, B.S.P. Mishra, S.K. Meher, Improved method of word embedding for efficient analysis of human sentiments, Multimedia Tools Appl. 79 (43) (2020) 32389–32413.

[647] A. Aue, M. Gamon, Customizing sentiment classifiers to new domains: A case study, in: Proceedings of Recent Advances in Natural Language Processing, Vol. 1, no. 3.1, Citeseer, 2005, pp. 1–2.

[648] L. Polanyi, A. Zaenen, Contextual valence shifters, in: Computing Attitude and Affect in Text: Theory and Applications, Springer, 2006, pp. 1–10.

[649] A. Das, S. Bandyopadhyay, Theme detection an exploration of opinion subjectivity, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1–6.

[650] A. Das, S. Bandyopadhyay, Subjectivity detection using genetic algorithm, Comput. Approach. Subject. Sent. Anal. (2010) 14–21.

[651] J.H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT Press, 1992.

[652] K. Sastry, D. Goldberg, G. Kendall, Genetic algorithms, in: Search Methodologies, Springer, 2005, pp. 97–125.

[653] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol. 63 (1) (2012) 163–173.

[654] S. Karimi, A. Shakery, A language-model-based approach for subjectivity detection, J. Inf. Sci. 43 (3) (2017) 356–377.

[655] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 111–119.

[656] Y. Belinkov, L. Màrquez, H. Sajjad, N. Durrani, F. Dalvi, J. Glass, Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 1–10.

[657] I. Chaturvedi, E. Cambria, S. Poria, R. Bajpai, Bayesian deep convolution belief networks for subjectivity detection, in: 2016 IEEE 16th International Conference on Data Mining Workshops, ICDMW, IEEE, 2016, pp. 916–923.

[658] N. Friedman, K.P. Murphy, S. Russell, Learning the structure of dynamic probabilistic networks, in: G.F. Cooper, S. Moral (Eds.), UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1998, pp. 139–147.

[659] A. Mogadala, V. Varma, Language independent sentence-level subjectivity analysis with feature selection, in: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, 2012, pp. 171–180.

[660] C. Largeron, C. Moulin, M. Géry, Entropy based feature selection for text categorization, in: Proceedings of the 2011 ACM Symposium on Applied Computing, 2011, pp. 924–928.

[661] S.-M. Kim, E. Hovy, Identifying and analyzing judgment opinions, in: Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 200–207.

[662] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 127–135.

[663] X. Wan, Co-training for cross-lingual sentiment classification, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language, 2009, pp. 235–243.

[664] C. Banea, R. Mihalcea, J. Wiebe, Multilingual sentiment and subjectivity analysis, Multiling. Nat. Lang. Process. 6 (2011) 1–19.

[665] I. Amini, S. Karimi, A. Shakery, Cross-lingual subjectivity detection for resource lean languages, in: Proceedings of the Tenth Workshop on Computational Approaches To Subjectivity, Sentiment and Social Media Analysis, 2019, pp. 81–90.

[666] I. Chaturvedi, E. Cambria, D. Vilares, Lyapunov filtering of objectivity for spanish sentiment model, in: 2016 International Joint Conference on Neural Networks, IEEE, 2016, pp. 4474–4481.

[667] A. Lopez, Statistical machine translation, ACM Comput. Surv. 40 (3) (2008) 1–49.

[668] K. Toutanvoa, C.D. Manning, Enriching the knowledge sources used in a maximum entropy part-of-speech tagger, in: 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000, pp. 63–70.

[669] A. Moro, F. Cecconi, R. Navigli, Multilingual word sense disambiguation and entity linking for everybody, in: Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272, 2014, pp. 25–28.

[670] N. Subrahmanya, Y.C. Shin, Sparse multiple kernel learning for signal processing applications, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2009) 788–798.

[671] Z. Zhang, Z.-N. Li, M.S. Drew, AdaMKL: A novel biconvex multiple kernel learning approach, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 2126–2129.

[672] G. Murray, G. Carenini, Predicting subjectivity in multimodal conversations, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 1348–1357.

[673] S. Raaijmakers, K.P. Truong, T. Wilson, Multimodal subjectivity analysis of multiparty conversation, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 466–474.

[674] B. Wrede, E. Shriberg, Spotting “hot spots” in meetings: human judgments and prosodic cues, in: Proceedings of EUROSPEECH, 2003, 2003, pp. 2805–2808.

[675] R. Banse, K.R. Scherer, Acoustic profiles in vocal emotion expression, J. Personal. Soc. Psychol. 70 (3) (1996) 614.

[676] R.E. Schapire, Y. Singer, BoosTexter: A boosting-based system for text categorization, Mach. Learn. 39 (2) (2000) 135–168.

[677] S. Raaijmakers, Sentiment classification with interpolated information diffusion kernels, in: Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising, 2007, pp. 34–39.

[678] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky, Linguistic models for analyzing and detecting biased language, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 1650–1659.

[679] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic inquiry and word count: LIWC 2001, Mahway: Lawrence Erlbaum Assoc. 71 (2001) (2001) 2001.

[680] D. Aleksandrova, F. Lareau, P.A. Ménard, Multilingual sentence-level bias detection in Wikipedia, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2019, pp. 42–51.

[681] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 427–431.

[682] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., Wiley, 2000.

[683] C. Hube, B. Fetahu, Neural based statement classification for biased language, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 195–203.

[684] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[685] K. Pant, T. Dadu, R. Mamidi, Towards detection of subjective bias using contextualized word embeddings, in: Companion Proceedings of the Web Conference 2020, 2020, pp. 75–76.

[686] N. Das, S. Sagnika, A subjectivity detection-based approach to sentiment analysis, in: Machine Learning and Information Processing, Springer, 2020, pp. 149–160.

[687] M. Bonzanini, M. Martinez-Alvarez, T. Roelleke, Opinion summarisation through sentence extraction: An investigation with movie reviews, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 1121–1122.

[688] A. Kamal, Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources, Int. J. Comput. Sci. Iss. 10 (5) (2013) 191.

[689] H.-C. Soong, N.B.A. Jalil, R.K. Ayyasamy, R. Akbar, The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques, in: 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics, 2019, pp. 272–277.

[690] G. Paltoglou, A. Giachanou, Opinion retrieval: Searching for opinions in social media, in: Professional Search in the Modern World, Springer, 2014, pp. 193–214.

[691] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 19–26.

[692] R. Cohen-Almagor, Fighting hate and bigotry on the Internet, Policy Internet 3 (3) (2011) 1–26.

[693] N.D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, Int. J. Multimed. Ubiquitous Eng. 10 (4) (2015) 215–230.

[694] B. Li, Y. Liu, A. Ram, E.V. Garcia, E. Agichtein, Exploring question subjectivity prediction in community QA, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 735–736.

[695] N. Aikawa, T. Sakai, H. Yamana, Community QA question classification: Is the asker looking for subjective answers or not? IPSJ Online Trans. 4 (2011) 160–168.

[696] V. Stoyanov, C. Cardie, J. Wiebe, Multi-perspective question answering using the Opqa corpus, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 923–930.

[697] M. Wan, J. McAuley, Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems, in: 2016 IEEE 16th International Conference on Data Mining, IEEE, 2016, pp. 489–498.

[698] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.

[699] D. Hillard, M. Ostendorf, E. Shriberg, Detection of agreement vs. disagreement in meetings: Training with unlabeled data, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–Short Papers-Volume 2, 2003, pp. 34–36.

[700] M. Galley, K. McKeown, J. Hirschberg, E. Shriberg, Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 669–677.

[701] D. Neiberg, K. Elenius, I. Karlsson, K. Laskowski, Emotion recognition in spontaneous speech, in: Proceedings of Fonetik, Citeseer, 2006, pp. 101–104.

[702] S. Somasundaran, J. Ruppenhofer, J. Wiebe, Detecting arguing and sentiment in meetings, in: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, 2007, pp. 26–34.

[703] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 2019, p. 6558.

[704] J. Zhang, B. Chen, L. Zhang, X. Ke, H. Ding, Neural, symbolic and neural-symbolic reasoning on knowledge graphs, AI Open 2 (2021) 14–35.

[705] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving Language Understanding by Generative Pre-Training, Technical Report, OpenAI, 2018.

[706] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, Technical Report, OpenAI, 2019.

[707] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[708] M. Ge, R. Mao, E. Cambria, A survey on computational metaphor processing techniques: From identification, interpretation, generation to application, Artif. Intell. Rev. (2023) http://dx.doi.org/10.1007/s10462-023-10564-7.

[709] E. Cambria, B. White, Jumping NLP curves: A review of natural language processing research, IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.

[710] R. Mao, G. Chen, X. Zhang, F. Guerin, E. Cambria, GPTEval: A survey on assessments of ChatGPT and GPT-4, 2023, arXiv:2308.12488.

[711] A. Cabrera, G. Neubig, Zeno chatbot report, 2023, GitHub repository, GitHub, https://github.com/zeno-ml/zeno-build/tree/main/tasks/chatbot/report. (Accessed: 18 May 2023).

[712] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M.T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, 2023, arXiv preprint arXiv:2303.12712.

[713] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst. 35 (2022) 24824–24837.