Artificial Intelligence Review (2023) 56:S1829–S1895
https://doi.org/10.1007/s10462-023-10564-7

A survey on computational metaphor processing techniques: from identification, interpretation, generation to application

Mengshi Ge[1] · Rui Mao[1] · Erik Cambria[1]

Published online: 11 August 2023

Abstract

Metaphors are figurative expressions frequently appearing daily. Given its significance in downstream natural language processing tasks such as machine translation and sentiment analysis, computational metaphor processing has led to an upsurge in the community. The progress of Artificial Intelligence has incentivized several technological tools and frameworks in this domain. This article aims to comprehensively summarize and categorize previous computational metaphor processing approaches regarding metaphor identification, interpretation, generation, and application. Though studies on metaphor identification have made significant progress, metaphor understanding, conceptual metaphor processing, and metaphor generation still need in-depth analysis. We hope to identify future directions for prospective researchers based on comparing the strengths and weaknesses of the previous works.

Keywords Metaphor identification · Metaphor interpretation · Metaphor generation · Conceptual metaphor processing · Metaphor processing application

1 Introduction

Metaphors are widely used figurative expressions in people's daily discourse. We may use metaphors much more frequently than we think. According to statistical corpus analysis (Steen et al. 2010b), metaphors appear in about a third of sentences in typical corpora. Experiments on machine translation by Mao et al. (2018) and sentiment analysis by Socher et al. (2013) showed that metaphors usually caused misunderstandings in those systems.

✉ Erik Cambria
cambria@ntu.edu.sg

Mengshi Ge
mengshi001@e.ntu.edu.sg

Rui Mao
rui.mao@ntu.edu.sg

[1] Continental-NTU Corporate Lab, Nanyang Technological University, Singapore, Singapore

Thus, metaphor processing is significant for natural language processing (NLP) downstream tasks.

Metaphors allow us to understand complex concepts, deliver abstract affective states, create diverse expressions, and frame human cognition. For example, love is an affection that emerges in human relationships. Metaphors give love more concrete meanings. Love is a *flame*[1], as it evokes a sense of passion. Love is a *roller-coaster ride*, offering the highest and lowest emotions during its period. Love is a *magnet*, which describes how two individuals are attracted to one another. These metaphors shape our understanding of love.

Metaphors frequently emerge in text, images, films, and music (Indurkhya 2013). Our survey only focuses on metaphors in text because linguistic researchers have laid a solid theoretical foundation for computational metaphor processing. Furthermore, vast amounts of textual data are readily accessible for people to collect and annotate for empirical studies. Finally, the theoretical findings about textual metaphors can facilitate metaphor processing in other modalities, such as metaphor processing from images (Fu et al. 2020; Zhang et al. 2021b). This survey does not discuss studies about visual metaphors due to our emphasis on natural language processing.

Metaphor studies on linguistics, psychology, and cognitive science (Turbayne 1964; Bickerton 1969; Billow 1975) started earlier than those in computational processes. Lakoff and Johnson (1980) proposed Conceptual Metaphor Theory (CMT) to explain metaphors as property transformations between two domains, usually from a more concrete domain (source domain) to a more abstract one (target domain). The mappings categorize metaphors and shape the way we think. Wilks (1975) presented that using metaphors involved deviating from the expected word choices based on selectional preferences. This can be reflected through word co-occurrence in current computational methods. Pragglejaz (2007) put forward Metaphor Identification Procedure (MIP) to annotate a large corpus in a standardized manner, focusing on the semantic contrast between the basic and contextual meaning of a unit. A considerable number of studies (Mason 2004; Choi et al. 2021; Qin and Zhao 2021; Ge et al. 2022) in computational metaphor processing benefited from the theories above. Meanwhile, studies from linguistics, psychology, and cognitive science (Prabhakaran et al. 2021; Han et al. 2022) also achieved exciting findings with the help of computational metaphor processing.

This survey introduces several sub-tasks of computational metaphor processing: metaphor identification, metaphor interpretation, conceptual metaphor processing, metaphor generation, and application. Metaphor identification aims to identify metaphoricity in a given text by different semantic units, e.g., words, phrases, and sentences. The objective of metaphor interpretation is to find literal expressions that deliver similar meanings to metaphors. Conceptual metaphor processing endeavors to identify and understand metaphors by generating and analyzing underlying source and target concepts. Metaphor generation aims to generate metaphors based on different syntactic patterns of metaphors. Metaphor applications employ computational metaphor processing models for other NLP downstream tasks. In the earlier stage, modeling ability was largely limited. Hence, metaphor identification datasets tend to deal with word pairs. Researchers used traditional machine learning methods to capture the semantic representation of metaphors, such as logic rules and clustering (Krishnakumaran and Zhu 2007). By 2018, word embedding models had developed considerably, attracting the attention of researchers. They focused on metaphors in sentences with specific syntactic patterns, such as verbs (Mao et al. 2018). In recent years, with the development of deep learning and the increase in computational ability, researchers have paid much attention to metaphors

[1] Italics are metaphors.

with more complex syntactic patterns (Steen et al. 2010a). Metaphor identification datasets with more diverse expressions of metaphors have become increasingly popular. Datasets for other sub-tasks are different due to various task formats. Metaphor interpretation datasets focus on providing explanations or literal expressions for metaphors. Conceptual metaphor processing datasets facilitate the presentation of the source and target concepts. Metaphor generation datasets include metaphors and their corresponding literal expressions in parallel.

To the best of our knowledge, the most recent survey papers about computational metaphor processing were from the studies of Rai and Chakraverty (2020) and Tong et al. (2021). Compared with their studies, our survey is distinctive in the following aspects: Rai and Chakraverty (2020) surveyed metaphor processing systems before 2019. However, with the pre-trained language models (PLM) upsurging in the NLP community, numerous PLM-based metaphor processing studies (Su et al. 2020a; Lin et al. 2021; Choi et al. 2021) were not included in their survey. More importantly, end-to-end metaphor processing techniques have made significant progress in the past years, greatly enhancing the usefulness of metaphor processing techniques in downstream applications and interdisciplinary research. For example, Mao et al. (2022a) proposed an end-to-end metaphor processing model (MetaPro) on all open-class words. The latest version (Mao et al. 2023) includes the functions such as metaphor identification, interpretation, and concept mapping. MetaPro enables a metaphor processing system to be conveniently applied in downstream tasks, e.g., sentiment analysis (Mao et al. 2022a) and depression detection (Han et al. 2022) in a data pre-processing fashion. Apart from metaphor identification, linguistic and conceptual metaphor interpretation, researchers have developed several studies on metaphor generation (Yu and Wan 2019; Chakrabarty et al. 2021; Stowe et al. 2021b) and application (Zheng et al. 2019; Cabot et al. 2020; Zhang et al. 2021a). However, Tong et al. (2021) did not cover academic progress in these aspects. Furthermore, compared with the work of Tong et al. (2021), our survey includes summaries of the tasks above with more explicit structures and more systematic comparisons in diverse dimensions.

The contribution of this survey is summarized below:

(1) We review and summarize the most recent works on computational metaphor processing with fine-grained technological trends and systematic comparison.
(2) We are the first to comprehensively review the current development of metaphor generation and application tasks, analyzing the main challenges, task definitions, and available solutions.
(3) We connect necessary theoretical research with advanced empirical studies and propose possible future directions.

In this survey, we focus on the latest algorithmic studies published in 2020-2022, notable and influential ones before 2020 , and studies with linguistic intuition. We do not explicitly discriminate metaphor from other figurative languages, such as metonymy, simile, sarcasm, pun, personification, and idiom, given that their boundaries are not clear-cut. The disagreement on this topic has yet to be solved (Burbules et al. 1989; Barcelona et al. 2000; Sam and Catrinel 2006). Moreover, frequently used corpus annotation guidelines (Lakoff 1994; Birke and Sarkar 2006; Pragglejaz 2007; Mohammad et al. 2016; Gutierrez et al. 2016) and recent metaphor processing studies (Martin 1990; Mason 2004; Tsvetkov et al. 2014; Shutova et al. 2016; Rei et al. 2017; Mao et al. 2018; Ge et al. 2022) similarly did not impose strict exclusion criteria on other figurative languages as well. Lastly, sarcasm detection, idiom detection, and pun detection are generally considered independent tasks (Hazarika et al. 2018; Li and Sporleder 2009; Ren et al. 2021). Distinguishing these figurative languages in our survey will lead to a massive extension of these tasks, which may not meet our research scope.

Thus, we follow the research scope commonly defined by previous computational metaphor processing methods and datasets, regarding "metaphor" as a general concept in this survey.

In the following sections, we first review theoretical research about metaphors (see Sect. 2) and available datasets for computational metaphor processing (see Sect. 3). Next, we analyze and summarize recent advanced studies of several sub-tasks by technical trends in the respective sections. Linguistic metaphor processing (see Sect. 4) studies metaphor identification (see Sect. 4.1) and interpretation (see Sect. 4.2) from the perspective of linguistic surface realization (Shutova 2015). Conceptual metaphor processing (see Sect. 5) focuses on generating or selecting concept mappings to elaborate metaphors. Metaphor generation (see Sect. 6) concentrates on creating metaphors from literal ones. Metaphor application (see Sect. 7) is to use metaphor processing techniques to support downstream tasks. Finally, we summarize future works in Sect. 8 and conclude this survey in Sect. 9.

2 Theoretical research

Metaphors in this survey are figurative expressions containing one or several words that produce semantic contrast between the basic and contextual meanings (Pragglejaz 2007). The contextual meaning is obtained under the whole contextual background, culture, sentiment, etc. The basic meaning is thought to be more concrete, body-related, and occurring earlier (Pragglejaz 2007). This definition is from a classical linguistic view (see Sect. 2.3), while another widespread theory, CMT (Lakoff and Johnson 1980), explored metaphors with cognitive concepts (see Sect. 2.1). Linguistic metaphors extend the manifestations of conceptual metaphors, while conceptual metaphors build the cognitive frame of linguistic metaphors, demonstrating the mutually complementary nature.

2.1 Conceptual metaphor theory

Lakoff and Johnson (1980) proposed CMT to explain metaphors from a cognitive perspective, abstracting linguistic metaphors with cross-domain concept mappings. They believed that instead of being a simple linguistic phenomenon, a metaphor revealed a distinct cognitive process. Based on the property or relation of concept mappings, metaphors produce more creativity and novelty in language understanding and expressions. CMT shows significant impacts on studies not limited to linguistics (Brinton and Brinton 2010; Barsalou 2019), psychological (Osbeck et al. 2010; Tileagă 2013), and computational (Mason 2004; Gagliano et al. 2016; Ge et al. 2022) fields.

(2.1) She *attacked* his argument.

In Example 2.1, the concept mapping in this metaphor is ARGUMENT IS WAR. This example transfers the aggression of WAR into ARGUMENT. Similar metaphors that fall into this category include:

(2.2) I have never *won* an argument with her.

(2.3) He *shot down* all of my arguments.

(2.4) Your claims are *indefensible*.

The above examples view ARGUMENT as WAR, so the person could have *won* an argument as they can win in a war, *shot down* arguments as they can shoot down enemies in a war, and become *indefensible*. Lakoff and Johnson (1980) pointed out that these war-related concepts sometimes emerge unconsciously in argument-related sentences. They demonstrated that people conceptually understand metaphors first before they generate linguistic metaphors in language. In this process, the transferred properties are not only aggression but also cooperation and strategy.

CMT categorizes metaphors into concept mappings between source and target domains. Lakoff and Johnson (1980) gave a general definition of source and target domains:

> "In a metaphor, there are two domains: the target domain, which is constituted by the immediate subject matter, and the source domain, in which important metaphorical reasoning takes place and that provides the source concepts used in that reasoning."

In Example 2.1, ARGUMENT is the target domain, commonly more abstract. WAR is the source domain, commonly more concrete. Moreover, Lakoff (1994) instantiated other source and target domains in common metaphors.

(2.5) LIFE IS A JOURNEY
Mary just *sails through* life.

(2.6) OPPORTUNITIES ARE OBJECTS
Seize the opportunity.

(2.7) DIFFICULTIES ARE CONTAINERS
We are *in* this thing together.

CMT proposed to use concept mappings to represent the cognition pattern of metaphors. However, the source and target concept representations were based on the personal judgment of the concept proposer while lacking a scientific definition. Appropriate concept representations and the abstraction level of concepts have yet to be discussed and standardized in CMT. In Example 2.1, the concept mapping ARGUMENT IS BATTLE is also acceptable. This could cause one to wonder if WAR or BATTLE might be more representative of the intended meaning. In line with this, CONFLICTION could also substitute these concepts. However, CONFLICTION might not be an appropriate concept to transfer the properties of WAR, e.g., cooperation, strategy, and aggression, compared to when WAR is used in concept mapping. This highlights the difficulty of using computational methods to model the concept mappings.

2.2 Selectional preference violation

Wilks (1975, 1978) developed a theory from a semantic perspective, termed Selectional Preference Violation (SPV), to explain and analyze metaphors. The author proposed that metaphors occurred when the selectional preferences of their context were broken. In other words, it indicates that metaphors contain word pair associations that fail to follow the frequent usage of some words. SPV is also applied in other NLP tasks such as word sense disambiguation (Gallant 1991; Agirre and Stevenson 2007), named entity recognition (Ratinov and Roth 2009), pronoun resolution (Bergsma et al. 2008), and textual inference (Ritter et al. 2010).

(2.8) My car *drinks* gasoline.

(2.9) My dog drinks water.

"drink" usually matches with an ANIMAL subject and a DRINKABLE_LIQUID object, such as "dog" and "water" in Example 2.9. However, in Example 2.8, "car" does not belong to the class of ANIMAL, and “gasoline” does not belong to the class of DRINKABLE_LIQUID. This causes the violation of selectional preference and yields a metaphor about “drink”. The word “drink” does not belong to the selectional preference of “car” or “gasoline”. Hence, SPV is bidirectional between the target word “drink” and the context, “car” and “gasoline”. Researchers (Mao et al. 2019; Choi et al. 2021; Su et al. 2021b) inspired by SPV usually compared the target word and the context information in model design.

SPV might not be proficient in detecting all types of metaphors, specifically conventional metaphors. “Conventional metaphors are metaphors that structure the ordinary conceptual system of our culture, which is reflected in our everyday language” (Lakoff and Johnson 1980). Novel metaphors are “not already part of the conceptual system of culture as reflected in its language and are capable of giving us a new understanding of our experience” (Lakoff and Johnson 1980). The public has widely accepted conventional metaphors, which exhibit almost equal frequency as literal expressions. These metaphorical meanings can be included in dictionaries as basic meanings (Sweetser 1990). In these cases, SPV may deliver weak signals in detecting conventional metaphors because conventional metaphors and their context may have achieved their selectional preference over time. In corpus-based and data-driven studies, one of the most common manifestations of SPV is the frequency of word co-occurrences. The frequency of conventional metaphors in a corpus can be larger than that of corresponding literal phrases. For example,

(2.10) She *spends* her time in reading novels.

The word collocations of <spend, `object`, time> occur the most frequently in the British National Corpus[2], surpassing many literal collocations such as <spend, `object`, money>.

Additionally, SPV can confuse metaphors with wrong collocations. For instance, “My car reads gasoline.” “read” violates the selectional preference of “car” and “gasoline”. However, this sentence is considered meaningless in language understanding and analysis. It can thus be considered as a wrong collocation. Thus, SPV may need to correct mistakes with wrong collocations.

2.3 Metaphor identification procedure

Pragglejaz (2007) presented the Metaphor Identification Procedure (MIP) to standardize the metaphor annotation process and combat the subjectivity issue in extensive corpus annotation. Steen et al. (2010a) put MIP into practice and proposed the annotation method from Metaphor Identification Procedure Vrije Universiteit (MIPVU). They annotated the largest token-level metaphor identification dataset, VU Amsterdam Metaphor Corpus (VUA). MIP resides in its role in providing a practical guideline that contributes to understanding what a metaphor is. Relative computational metaphor processing studies also proved the effectiveness of the mechanism of MIP in guiding model design (Song et al. 2021; Lin et al. 2021; Ottolina et al. 2021; Qin and Zhao 2021). MIP takes the following steps (Pragglejaz 2007):

(1) Read the entire text-discourse to establish a general understanding of the meaning.
(2) Determine the lexical units in the text-discourse.
(3) (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.

[2] http://www.natcorp.ox.ac.uk/.

(b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be

(i) More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.

(ii) Related to bodily action.

(iii) More precise (as opposed to vague)

(iv) Historically older.

Basic meanings are not necessarily the most frequent meanings of the lexical unit.

(c) If the lexical unit has a more basic current-contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.

(4) If yes, mark the lexical unit as metaphorical.

(2.11) Fear *clogged* his mind.

Here, we refer to Example 2.11 to demonstrate how to conduct MIP. In Step 1, we understand the general meaning of the sentence, “He was quite scared.” In Step 2, the lexical units in this sentence are “fear/clog/his/mind”. In Step 3, we compare the contextual meaning and basic meaning of each lexical unit. We look up the basic meaning for each unit from Macmillan English Dictionary for Advanced Learners (Rundell and Fox 2002), which was the recommended dictionary for MIP.

fear : The contextual meaning is a scared emotion from the male figure. According to the dictionary, the basic meaning is a feeling of a person when he is frightened. The contextual and basic meanings are similar, so “fear” is literal.

clog : The contextual meaning is that the fear is so unbearable that his mind fails to work as usual. In this situation, “fear” is compared to an object, and the mind is compared to a pipe or passage that something can go through. The word “clogged” concretes abstract words, namely “fear” and “mind”, into more concrete concepts. On the other hand, the basic meaning of “clog” is to block a pipe or passage. The contextual and basic meanings are different. Thus, “clog” is metaphorical.

his : The contextual meaning is that the emotion of fear is from a male figure. According to the dictionary, the basic meaning is defined to show that something belongs to a male figure who has been mentioned before. The contextual and basic meanings are similar, so “his” is literal.

mind : The contextual meaning is where fear or other feelings exist. The basic meaning is “the part of a person that thinks, knows, remembers, and feels things” in the dictionary. The contextual and basic meanings are similar, so “mind” is literal.

In general, “clogged” in this sentence is metaphorical, while others are literal. The identification process is analyzed at the token level and conducted in a pipeline style. It can be done step by step, and discussion can happen at each step where disagreements exist. To some degree, MIP is in line with SPV. Comparing basic and contextual meanings is similar to analyzing the violation of selectional preference between a metaphor and its context. Both MIP and SPV have explained metaphors from linguistic aspects. Furthermore, MIP is more practical for metaphor annotation than CMT because it can address the annotation process without requiring concepts and concept mappings.

(2.12) I *see* your points. We can discuss them later.

(2.13) However I say ultimately, because being and staying focused on one subject means always to *discard* other subjects. (Klebanov et al. 2018)

However, the limitations of MIP also exist. The boundary between basic and contextual meanings sometimes needs to be clarified. For a word with multiple meanings, it is particularly challenging to distinguish which one is the basic meaning. For example, "see" has two ordinary senses in English Oxford Dictionary: "perceive with the eyes; discern visually" and "discern or deduce after reflection or from information; understand." It has yet to be agreed upon whether the latter is one of the basic meanings (Sweetser 1990) (see Example 2.12). This leads to high subjectivity when determining the basic and contextual meanings. Thus, the subjectivity of annotators can influence the metaphor annotation results. The metaphoricity annotation of *discard* in Example 2.13 similarly reflected this issue. According to its dictionary meaning, "to get rid of something that you no longer want or need", the annotation label depends on whether "subjects" refer to concrete objects with visible shapes or abstract terms related to the non-physical world in annotators' thoughts. Shutova and Teufel (2010) made annotations by MIP procedures. The average of inter-annotator agreement among the three annotators is 0.64. The low score proves the ambiguity of metaphor annotations for humans.

3 Datasets

3.1 Metaphor identification datasets

Birke and Sarkar (2006) developed a dataset, TroFi, containing 50 target verbs whose metaphoricity would be identified in this dataset. They collected 3737 sentences from Wall Street Journal Corpus. The dataset was annotated by an unsupervised word sense disambiguation-based clustering algorithm and evaluated by two authors. Metaphors arose 43.5% of the annotated sentences.

Steen et al. (2010b) annotated the metaphoricity of each token in sentences, forming the largest all-word annotated metaphor corpus, VU Amsterdam Metaphor Corpus (VUA[3]), with the guidance of MIP. The sentences were collected from British National Corpus (BNC) (Consortium 2007) in four genres: news, academic, fiction, and conversation. The corpus contains 10,567 sentences, of which 11.6% have metaphors. The corpus also labeled metaphor types: indirect metaphors, direct metaphors, implicit metaphors, and borderline metaphors. Indirect metaphors (see Example 3.1), the largest group in this corpus, show the contrast between the contextual and basic meanings. Direct metaphors (see Example 3.2) compare the expression directly via language use. An implicit metaphor (see Example 3.3) refers to an underlying connection in the discourse, which indicates a metaphorical concept. Borderline metaphors (see Example 3.4) refer to unconfident annotations due to ambiguous context or disagreement among annotators. Examples 3.1–3.4 are from Steen et al. (2010b):

(3.1) Professional religious education teachers like Marjorie B Clark (Points of View, today) are doing *valuable* work in many secondary schools ...

(3.2) ... he's like a *ferret*.

(3.3) Naturally, to embark on such as step is not necessarily to succeed in realizing *it*.

[3] http://www.vismet.org/metcor/documentation/home.html.

(3.4) But by the time I had turned off the road from Bellingham at Kielder village and driven *up* the bumpy Forest Drive to East Kielder Farm ...

Tsvetkov et al. (2014) released an adjective-noun word pair dataset (TSV[4]). The training set was collected from public resources by two annotators. Additional annotators examined the sentences, abandoning duplicates, weak metaphors, and metaphorical phrases. The sentences in the test set were searched in TenTenWeb corpus[5], containing words frequently associated with the 1000 most commonly used adjectives. The training set contains 884 metaphorical and 884 literal sentences, whereas the test set contains 100 metaphorical and 100 literal sentences.

Mohammad et al. (2016) developed a dataset (MOH-X[6]) by selecting verbs with three to ten senses and their corresponding instantiated sentences from WordNet. Each label was annotated by ten annotators from the crowd-sourcing platform CrowdFlower[7]. At least 70% of annotators reached a consensus on the labeled data. The dataset has 1639 (1230 literal expressions and 409 metaphors) sentences and 440 target verbs.

Shutova et al. (2016) extracted verb-object and verb-subject word pairs from the sentences in MOH-X, eliminating those that were pronominal or clausal. The dataset (MOH) contains 647 verb-noun pairs, 316 metaphorical and 331 literal.

Gutierrez et al. (2016) developed a dataset (GUT[8]) focusing on 23 adjectives with both metaphorical and literal meanings. The selected 8592 word pairs (3991 literal, 4601 metaphorical) occurred more than ten times in their corpora (2011 dump of English Wikipedia, the UKWaC (Baroni et al. 2009), BNC (Consortium 2007), and the English Gigaword corpus (Graff and Cieri 2003)).

Klebanov et al. (2018) collected 240 essays from the ETS Corpus of Non-Native Written English[9] and developed a metaphor dataset (TOEFL[10]). An English instructor and the lead author of this paper did the annotation. At least one annotator agreed upon an annotated metaphor in the dataset. The training set contains 180 essays and 2,741 sentences. The test set contains 60 essays and 968 sentences.

Zayed et al. (2019) presented a crowd-sourcing approach for building a metaphor identification dataset and released one (ZayTw) sourced from Twitter on general and political topics. They applied a weakly supervised classifier (Zayed et al. 2018) on the source data and filtered data with criteria such as verb balance, sense coverage, and size before crowd-sourcing.

Given the recent multimodal and multi-task learning trend, Zhang et al. (2021b) provided a multimodal dataset from social media and advertisements. In addition to a metaphor label, each data record contains an image, source and target concepts, sentiment, and author intent labels. The authors followed Tasić and Stamenković (2015) and divided data records into three categories. They are text dominant, image dominant, and complementary, indicating whether text or image expresses metaphoricity. This dataset can be used for conceptual metaphor processing, metaphor identification, sentiment analysis, and intent prediction.

[4] https://github.com/ytsvetko/metaphor.

[5] http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen.

[6] http://saifmohammad.com/WebPages/metaphor.html.

[7] www.crowdflower.com.

[8] http://bit.ly/1TQ5czN.

[9] https://catalog.ldc.upenn.edu/LDC2014T06.

[10] https://github.com/EducationalTestingService/metaphor.

Xu et al. (2022) released a public meme dataset, MET-Meme[11], for meme studies in cooperation with metaphorical information. Following Tasić and Stamenković (2015), the authors also categorized data records into text dominant, image dominant, and complementary. NLP postgraduate students and research assistants were responsible for metaphoricity annotations, source and target concepts. A professional crowd-sourcing company completed the annotations for sentiment, offensive, and intention labels. The inter-annotator agreement for metaphor annotation was not provided. The image data were from Twitter, Google, MEMOTION (Sharma et al. 2020), Weibo, and Baidu images. The textual data were obtained by OCR API (Sivakumar et al. 2018). The paper provided baseline results of four sub-tasks: metaphor detection, sentiment analysis, intention detection, and offensiveness detection.

3.2 Metaphor interpretation datasets

Bizzoni and Lappin (2018) annotated a dataset[12] for paraphrasing metaphors. There are 200 sets of sentences, each containing a metaphorical sentence and four literal sentences: a strong paraphrase, a loose paraphrase, and two non-paraphrases of the metaphorical sentence. The dataset was manually developed with a relatively small size.

Zayed et al. (2020b) categorized metaphors into three types (lexical substitution, paraphrase generation, and definition generation) and released a dataset for verb metaphor definition generation. The author believed that when a new language learner found a verbal metaphor hard to interpret, the learner might look up a dictionary to understand the meanings of the verb. The definitions of verbal metaphors were sourced from idiomatic or metaphorical senses in a dictionary. The Amazon Mechanical Turk (AMT) annotators were asked to choose the most appropriate definition from three candidates or provide one themselves.

Liu et al. (2022) presented a task inspired by the Winograd schema (Levesque et al. 2012) to explore the power of language models in metaphor interpretation. The task was to make language models to choose or generate implications for two metaphorical sentences with opposite meanings. They built a corresponding interpretation dataset, termed Fig-QA[13]. Fig-QA contains 10,256 rare or creative metaphors, where workers on AMT generated the interpretations.

3.3 Concept mapping datasets

Lakoff (1994) first attempted to create a comprehensive knowledge base for source and target concept mappings, termed Master Metaphor List (MML). It includes 791 nested concept mappings with corresponding metaphorical examples. MML remains a draft status that cannot be utilized conveniently in the study. Some researchers criticized it for the unclear structure of concept mappings (Lönneker-Rodman 2008), confused taxonomies, and non-exclusive concept classes (Shutova and Teufel 2010).

Shutova and Teufel (2010) proposed a procedure for concept mapping annotations. They extracted a subset from BNC and provided source and target concept lists for three native English speakers. The annotators had a linguistic background. However, the process of annotations remained challenging in determining the abstraction level.

[11] https://github.com/liaolianfoka/MET-Meme-A-Multi-modal-Meme-Dataset-Rich-in-Metaphors.

[12] https://github.com/yuri-bizzoni/Metaphor-Paraphrase.

[13] https://github.com/nightingal3/Fig-QA.

Aiming at enterprise-level development and evaluation of metaphor identification, Mohler et al. (2016) released a large-scale word pair metaphor dataset in four languages (English, Spanish, Russian, and Farsi). The annotation contained metaphoricity rates on a 4-point scale (clear, conventionalized, weak metaphor, and literal), source and target concept domains, affect polarity, and intensity rates. The multi-lingual data were sourced from ClueWeb09 corpus[14] and Debate Politics Online Forum[15] for English, Spanish Gigaword Corpus (Consortium 2011) for Mexican Spanish, RuWac Corpus[16] for Russian, and Hamshahri Corpus[17] of Iranian newswire texts for Farsi. This dataset (LCC[18]) contains over 40,000 data points for each language.

3.4 Metaphor generation datasets

Chakrabarty et al. (2021) created a large-scale parallel corpus (MGen[19]) for metaphor generation training. They extracted metaphorical sentences from Gutenberg Poetry Corpus[20] (Jacobs 2018) by a BERT-based classifier trained with VUA. A masked language model was employed to generate literal counterpart candidates. They utilized COMET (Bosselut et al. 2019) to ensure that the semantic meanings of metaphorical and literal parts were similar.

Li et al. (2022) publicly released a Chinese nominal metaphor dataset for generation and identification tasks. The 6,257 data points from children's literature, Chinese literature, and translated literature were annotated by three native Chinese annotators with a 0.84 agreement rate. Each data point indicates a sentence, source and target concepts, and a comparator.

3.5 Summary

Table 1 shows the Parts of Speech (PoS) of metaphors in the datasets above. The "All" column contains close-class words besides open-class (e.g., verb, adjective, adverb, noun) words. Verbal metaphor is the most prevalent metaphor class in these datasets. This partly influences the choice of PoS when building models. Noun and adjective metaphors are also included in multiple datasets. Few datasets annotated adverbial metaphors because they are less common than other PoS. Additionally, annotating metaphors for multi-word expressions (MWE) and all PoS is unusual in the community. However, this is important because MWE is a common linguistic phenomenon in figurative expressions, possibly appearing in idioms, similes, and personifications.

Table 2 shows the data format for each dataset. Sentence-level means the label denotes the metaphoricity of a complete sentence. Relation-level means the label is for a dependent word pair. Token-level means the label indicates the metaphoricity of a word in a sequence. In our context, concept mapping denotes the source and target concepts directly extracted from a sentence without abstraction. Concept domains denote abstracted concept categories for source and target domains (see details in Sect. 2.1). Most datasets are released with sentences, reflecting the importance of contextual information. The data format largely depends on the

[14] http://www.lemurproject.org/clueweb09.php/.

[15] http://www.debatepolitics.com/.

[16] http://corpus.leeds.ac.uk/tools/ru/ruwac-parsed.out.xz.

[17] http://ece.ut.ac.ir/dbrg/hamshahri/.

[18] http://www.languagecomputer.com/metaphor-data.html.

[19] https://github.com/tuhinjubcse/MetaphorGenNAACL2021.

[20] https://github.com/aparrish/gutenberg-poetry-corpus.

Table 1 Metaphor datasets by author and part-of-speech

Task	Paper	Verb	Noun	Adjective	Adverb	MWE	All
Id.	Birke and Sarkar (2006)	✓					
	Steen et al. (2010b)						✓
	Tsvetkov et al. (2014)			✓			
	Gutierrez et al. (2016)			✓			
	Mohammad et al. (2016)	✓					
	Shutova et al. (2016)	✓					
	Klebanov et al. (2018)						✓
	Zayed et al. (2019)	✓	✓				
	Zhang et al. (2021b)	✓	✓				
	Xu et al. (2022)	✓	✓	✓			
INTPN	Bizzoni and Lappin (2018)	✓	✓	✓		✓	
	Zayed et al. (2020b)	✓					
	Liu et al. (2022)						✓
CMP	Lakoff (1994)	✓	✓	✓			
	Shutova and Teufel (2010)	✓					
	Mohler et al. (2016)	✓	✓	✓	✓	✓	
Gen.	Chakrabarty et al. (2021)	✓					
	Li et al. (2022)		✓				

MWE multi-word expression, *Id* metaphor identification, *INTPN* metaphor interpretation, *CMP* conceptual metaphor processing, *Gen.* metaphor generation

task. For the identification task, token-level annotations are the most common label format. Zhang et al. (2021b), Xu et al. (2022) delivered a multimodal dataset. Thus, they additionally included image and concept modalities to show the metaphoricity of the images and text in the datasets. The data formats in the interpretation task are more diverse than other tasks since the community has yet to reach a consensus on the task format. Researchers have tried multiple formats to perform this task. All the conceptual metaphor datasets contain concept domains, which are the crux of conceptual metaphor processing. A few researchers (Mohler et al. 2016; Zhang et al. 2021b; Xu et al. 2022) have noticed the connections between metaphors and sentiments, including sentiment, intent, and affection labels.

Table 3 shows the sources and genres of the datasets, where the features distribute sparsely. Non-domain-specific data sources, such as BNC and Wikipedia, are relatively popular. This demonstrates that domain-specific datasets are not preferred in the community because metaphors are likely to appear in different text types. Social media, such as Twitter, is another popular data source because its language is comparatively more colloquial and figurative. Metaphor generation datasets tend to source data from literature and poetry because these genres contain richer metaphors in a document.

Table 4 provides the basic information about dataset acquisition, languages, and size. Most of them are publicly available online or obtainable by contacting authors. At present, most datasets focus on English metaphors. There are relatively few datasets in other languages (Tsvetkov et al. 2014; Mohler et al. 2016; Xu et al. 2022). Cultural differences lead to different metaphors in different languages. Therefore, multi-lingual annotated datasets are of great significance for metaphor research. However, these studies did not mention the differences between monolingual, bilingual, and multi-lingual dataset annotation.

Table 2 Metaphor datasets by author and data format

Task	Paper	Sent.-level	Rel.-level	Token-level	Word pair	Sent.	Concept domain	Concept mapping	Image	Concept modality	Stmt.	Int.	Others
Id.	Birke and Sarkar (2006)			✓		✓							
	Steen et al. (2010b)			✓		✓							Metaphor type
	Tsvetkov et al. (2014)		✓		✓								
	Gutierrez et al. (2016)		✓		✓								Occurence count
	Mohammad et al. (2016)			✓		✓							metaphoricity rating
	Shutova et al. (2016)		✓		✓								Relation
	Klebanov et al. (2018)			✓		✓							
	Zayed et al. (2019)			✓		✓							
	Zhang et al. (2021b)			✓		✓		✓	✓	✓	✓	✓	Data resource
	Xu et al. (2022)			✓		✓		✓	✓	✓	✓	✓	Offensiveness degree
INTPN	Bizzoni and Lappin (2018)	✓				✓							1 metaphorical sentence and 4 literal candidates with scores
	Zayed et al. (2020b)		✓		✓								Definition
	Liu et al. (2022)	✓				✓							Paired sentences with opposite meanings and interpretations
CMP	Lakoff (1994)	✓				✓	✓						
	Shutova and Teufel (2010)			✓		✓	✓						
	Mohler et al. (2016)		✓		✓	✓	✓						Affective polarities, and intensity rating
Gen.	Chakrabarty et al. (2021)			✓		✓							
	Li et al. (2022)	✓				✓							

Sent. sentence, *Rel.* relation, *Stmt.* sentiment, *Int.* intention, *The fourth column (Sent.)* whether a dataset contains sentential information

Table 3 Metaphor datasets by author, data source, and genre

Task	Paper	NDS	Jou.	Lit.	Poetry	Essay	SM	Web	KB	Ad.	Pol.	MC
Id.	Birke and Sarkar (2006)		✓									
	Steen et al. (2010b)	✓										
	Tsvetkov et al. (2014)		✓					✓				
	Gutierrez et al. (2016)	✓										
	Mohammad et al. (2016)								✓			
	Shutova et al. (2016)								✓			
	Klebanov et al. (2018)					✓						
	Zayed et al. (2019)						✓					
	Zhang et al. (2021b)						✓			✓		
	Xu et al. (2022)						✓	✓				
INTPN	Bizzoni and Lappin (2018)											✓
	Zayed et al. (2020b)						✓		✓			
	Liu et al. (2022)											✓
CMP	Lakoff (1994)					✓						
	Shutova and Teufel (2010)	✓										
	Mohler et al. (2016)	✓									✓	
Gen.	Chakrabarty et al. (2021)				✓							
	Li et al. (2022)			✓								

NDS non-domain-specific data sources, *Jou.* journal, *Lit.* literature, *SM* social media, *KB* knowledge base, *Ad.* advertisement, *Pol.* politics, *MC* manually created

Table 4 Metaphor datasets by author and basic information

Task	Paper	Public online	Request	Partly avail.	EN	ZH	Others	SZ
Id.	Birke and Sarkar (2006)	✓			✓			3737
	Steen et al. (2010b)	✓			✓			16,204
	Tsvetkov et al. (2014)	✓			✓		ES, RU, FA	1990
	Gutierrez et al. (2016)	✓			✓			8592
	Mohammad et al. (2016)	✓			✓			1639
	Shutova et al. (2016)		✓		✓			647
	Klebanov et al. (2018)	✓			✓			3710
	Zayed et al. (2019)		✓		✓			2500
	Zhang et al. (2021b)		✓		✓			10,437
	Xu et al. (2022)	✓			✓	✓		10,045
INTPN	Bizzoni and Lappin (2018)	✓			✓			200
	Zayed et al. (2020b)		✓		✓			1500
	Liu et al. (2022)	✓			✓			10,256
CMP	Lakoff (1994)	✓			✓			791
	Shutova and Teufel (2010)		✓		✓			761
	Mohler et al. (2016)		✓	✓	✓		ES, RU, FA	80,100
Gen.	Chakrabarty et al. (2021)	✓			✓			93,498
	Li et al. (2022)	✓				✓		6257

EN English, *ZH* Chinese, *RU* Russian, *ES* Spanish, *FA* Farsi, *SZ* data size

Table 5 shows the annotation information of different datasets. Most annotation work was finished by in-house students, research assistants, and crowd-sourcing platforms, such as CrowdFlower and AMT. About half of the datasets were annotated by 2-5 annotators. The datasets from Steen et al. (2010b) and Bizzoni and Lappin (2018) achieved relatively high annotation agreement.

Most published datasets showed metaphor annotation in a binary format (metaphorical/literal). Steen et al. (2010a); Mohler et al. (2016) annotated metaphors on a more fine-grained scale. Steen et al. (2010b) additionally annotated indirect, direct, implicit, and borderline metaphors. Mohler et al. (2016) supplemented categories for weak and conventional metaphors. These resources provided a foundation for more in-depth computational metaphor processing.

The most common annotation method was manual labeling, where annotators were asked to choose a correct answer from given choices. Annotators were also encouraged to propose an answer if none of the given choices were suitable in the studies involving cognitive understanding (Shutova and Teufel 2010; Zayed et al. 2020b). Liu et al. (2022) instructed annotators to generate creative metaphors, which gave annotators the most extensive freedom. Some researchers applied machine learning or deep learning classifiers to reduce annotation workload. Birke and Sarkar (2006) used an unsupervised word sense disambiguation-based clustering algorithm to annotate labels. Zayed et al. (2019); Mohler et al. (2016) applied metaphor identification models to obtain weakly-annotated datasets. Chakrabarty et al. (2021) utilized a metaphor classifier to extract metaphors and masked metaphorical words to generate literal expressions by BERT.

To better annotate metaphors and control the quality of annotation, all the studies mentioned that authors chose annotators with related backgrounds and gave annotators clear instructions and explicit examples. Most datasets contain data records with more than one annotator completing the annotation. Shutova and Teufel (2010); Klebanov et al. (2018); Zayed et al. (2019, 2020b) set a training session before the annotation. Steen et al. (2010b); Zhang et al. (2021b); Xu et al. (2022) held regular meetings during the annotation process to discuss annotation problems. Zayed et al. (2019); Liu et al. (2022) tested crowd-sourcing annotators before the formal annotation process and selected top-performing annotators. Mohammad et al. (2016); Xu et al. (2022) discarded data records with IAA below a set threshold. Birke and Sarkar (2006); Liu et al. (2022) evaluated the quality after the annotation by classifiers or annotators. Furthermore, Klebanov et al. (2018) managed the second annotation for data with low IAA and finally assigned metaphor labels to words agreed upon by at least one annotator. Steen et al. (2010b) allowed annotators to comment on others' annotations. Any data record (Xu et al. 2022) could be discarded if any annotator strongly disagreed with the result.

Different tasks have different requirements for datasets. Firstly, the dataset should be relevant to the problems that researchers focus on. Lakoff (1994); Shutova and Teufel (2010); Mohler et al. (2016) contain concept metaphors that can be considered in conceptual metaphor processing. The dataset size has an important influence on the generalization of models. Larger datasets (Steen et al. 2010b; Mohler et al. 2016; Chakrabarty et al. 2021) can support training more complicated models. However, some researchers were more interested in metaphors from a specific domain. For example, Zayed et al. (2019) focused on social media data, while Chakrabarty et al. (2021) collected poetry data. Furthermore, annotation quality and availability are essential factors when choosing a dataset.

Table 5 Metaphor datasets by author and annotation information. RCT denotes recruited

Task	Paper	In-house	Volunteer	RCT	Crowd -source	Annotator	IAA
Id.	Birke and Sarkar (2006)	✓				2	$\kappa = 0.77$
	Steen et al. (2010b)	✓				5	$FK = 0.85$
	Tsvetkov et al. (2014)					5	$FK = 0.80$
	Gutierrez et al. (2016)	✓	✓			2	$FK = 0.80$
	Mohammad et al. (2016)				✓	> 10	
	Shutova et al. (2016)					NA	
	Klebanov et al. (2018)	✓		✓		2	$\kappa = 0.62$
	Zayed et al. (2019)				✓	5	$FK > 0.7$
	Zhang et al. (2021b)	✓			✓	> 15	$FK = 0.67$
	Xu et al. (2022)			✓	✓	> 20	
INTPN	Bizzoni and Lappin (2018)	✓			✓	> 20	$PC = 0.93$
	Zayed et al. (2020b)				✓	6	$FK = 0.48$
	Liu et al. (2022)				✓	NM	
CMP	Lakoff (1994)	✓				NM	
	Shutova and Teufel (2010)		✓			3	$\kappa = 0.61$
	Mohler et al. (2016)	✓				9	> 0.86
Gen.	Chakrabarty et al. (2021)					NA	
	Li et al. (2022)			✓		3	$KA = 0.84$

IAA denotes inter-annotator agreement, *NA* means the methods in that paper did not need annotators, *NM* means the authors did not mention the number of annotators, κ Cohen's kappa (Cohen 1960), *FK* Fleiss' kappa (Fleiss 1971), *PC* Pearson correlation, *KA* denotes Krippen-dorff's alpha (Krippendorff 2011)

4 Linguistic Metaphor Processing

Following Shutova (2015), linguistic metaphor processing focuses on the manifestations of metaphors in text, which involves two main tasks: metaphor identification and metaphor interpretation. These two tasks are the most widely studied research topics in computational metaphor processing (Martin 1990; Shutova 2010; Mohler et al. 2013; Rai et al. 2019; Su et al. 2020a).

(4.1) The sky is *crying*. (Indurkhya 2013)

We use Example 4.1 to demonstrate how we identify and understand metaphors linguistically. The word “crying” is usually used to express the shedding of tears when one is feeling sad or emotional. The subjects of the sentences are usually human beings. This sentence highlights the similarity of the sky raining and a human being crying, consequently comparing the sky to a human being. Literally, the sky cannot cry, so “crying” is metaphorical. It means that it is raining.

(4.2) When I see the picture of her face with two big pale blue eyes, I see the *sky* is crying. (Indurkhya 2013)

However, the identification and interpretation of metaphors depend on contextual information. Example 4.2 shows the same sentence as Example 4.1 with extra contextual information. The words “sky” and “big pale blue eyes” share the same color and have the similar property of raining and crying. Here, the “sky” is not a physical object. It is a metaphorical concept that is compared to eyes that are crying. Comparing Example 4.1 with Example 4.2, we can conclude that the same sentence can be marked with different metaphorical labels and meanings, given different contexts.

(4.3) I don’t think this relationship is *going anywhere*. (Kovecses 2010).

(4.4) This young man knows how to *climb* the social *ladder* (Mohammad et al. 2016).

A metaphor can be expressed via a single word, such as Examples 2.1 and 2.8, or several words, such as Examples 4.3 and 4.4. “go anywhere” and “climb the ladder” are literal expressions. Once they are read in their contexts, semantic contrasts exist between the basic and contextual meanings. In Example 4.3, the contextual meaning of “go anywhere” is to make progress. This type of metaphor is commonly called a metaphorical MWE. Formally, a metaphorical MWE can be defined as a metaphor consisting of multiple metaphorical words, whereas these words can be literally used in an independent context.

Another particular type of metaphor is the extended metaphor. It refers to sequentially used metaphors under the same concept frame at the discourse level. Example 4.5 is a typical extended metaphor written by William Shakespeare in *As you like it* (Act 2, Scene 7) (Shakespeare 2019).

(4.5) All the world’s a *stage*, / And all the men and women merely *players*: / They have their *exits* and their *entrances*; / And one man in his time *plays* many *parts*, [..]

This speech starts by comparing the “world” to the “stage”. Under this scenario, “men” and “women” are compared to “players” as an extension of “stage”. All the metaphors in this example are from the same domain. The concept projections of source and target concepts can extend to the whole scene in the act. Thus, extended metaphors can help understand multiple complex concepts in discourse with parallel comparisons.

Understanding metaphors needs commonsense knowledge, such as life experience or cultural information. The writers and the readers of metaphors resemble encoders and decoders. These two should share the same background information to make sure that the metaphors expressed by writers can be identified and understood correctly by the readers.

(4.6) The West Lake is *XiShi*. She is suitable for both light makeup and heavy makeup (Su et al. 2020b).

Example 4.6 compares the West Lake to XiShi (an ancient Chinese beauty) to convey the beautiful scenery. XiShi is a culture-specific concept. One unfamiliar with Chinese history cannot identify or understand this metaphor.

4.1 Metaphor identification

Metaphor identification means identifying a target's metaphoricity as either a metaphor or a literal expression. The task can be categorized as sentence-level, relation-level, and token-level metaphor identification sub-tasks.

(1) The sentence-level task takes a sentence as input and outputs a label, indicating whether the sentence has a metaphorical meaning.
(2) The relation-level task takes a dependent word pair as input, which usually targets subject-verb, verb-direct object, or adjective-noun dependent relationships.
(3) The token-level tasks can be categorized as sequence labeling (SEQ) and classification (CLS) tasks (Gao et al. 2018). The sequence-labeling task inputs a sentence and outputs a label sequence, indicating the metaphoricity of each token in the sentence. The classification task inputs a sentence with a known target word and then outputs a metaphoricity label for the target word.

4.1.1 Sentence-level metaphor identification

Krishnakumaran and Zhu (2007) focused on processing noun metaphors and categorized them into three classes, nominal metaphor: subject-verb-object metaphor, and adjective-noun metaphor. They utilized hyponym relations in WordNet (Miller 1998) and bigram counts in Web 1T corpus (Brants and Franz 2006) to seek possible relations between entities. This work relied intensely on external resources and could not deal with polysemous issues.

Tsvetkov et al. (2013) first studied cross-lingual metaphor detection. Since not all languages have resources as rich as those found in English while semantic features can maintain across languages, they proposed a semantic feature-based classifier to detect metaphors without extensive lexical resources. The method concatenated semantic categories, abstractness degrees, and named entity types and achieved 0.76 and 0.78 F1 scores on Russian and English datasets, respectively.

Mohler et al. (2013) proposed a domain-specific classifier with semantic signatures to detect metaphors in unstructured texts. They employed WordNet and Wikipedia to explore word senses and related documents to construct domain signatures by clustering. Subsequently, they compared the signatures of known metaphors and candidates by five hand-coded metrics. The classifiers, processing five metrics, were based on machine learning tools. The performance showed that metaphorical concepts could share common semantic signatures in a specific domain.

The modality norm describes every word in terms of six primary senses (auditory, gustatory, haptic, visual, olfactory, and interoceptive). Given the hypothesis that a metaphor showed a shift in modality from source to target concepts, Wan et al. (2020) concatenated the modality norm[21] with word embeddings (GloVe (Pennington et al. 2014)). The results outperformed several BERT-based baselines, showing that the modality norm is a helpful feature in identifying metaphors.

4.1.2 Relation-level metaphor identification

Tsvetkov et al. (2014) demonstrated that lexical semantic features were reliable in detecting metaphor patterns in word pairs by feeding three feature categories (abstractness and imageability, supersenses[22], and vector-space word representations) into a random forest model. They trained a logistic regression classifier to output scores using vector-space word representations as features for words without abstractness and imageability scores. It had promising performance on subject-verb-object and adjective-noun metaphor datasets.

When a person comprehends metaphors, information from other modalities also plays a role. Among them, visual features are the most intuitive supplement to the text. Shutova et al. (2016) first attempted to combine word embeddings and visual embeddings in metaphor identification. The visual embeddings were obtained by a deep convolutional neural network (CNN) processing ten images from Google Images per word.

Bulat et al. (2017) aimed to test the hypothesis that attribute-based semantic representations, e.g., property norm, could perform better than dense linguistic representations, e.g., Word2Vec (Mikolov et al. 2013). Their experiments compared a model with semantic and linguistic representations and a model with linguistic representations, demonstrating that semantic representations could capture extra information besides linguistic representations. This work first utilized large-scale attribute-based semantic representations on metaphor identification and next brought the community a new perspective on semantic and linguistic representations. However, this work was confined to finding an attribute with an appropriate abstraction level for the representation.

Rei et al. (2017) applied neural networks to capture the semantic information of metaphorical and literal texts at the relation level. They used a neural network layer to formulate the calculation of cosine similarity. The embedding of a target word and its word pair were input into the model. The experiments showed that this similarity-based network outperformed a simple feed-forward neural network (FNN[23]).

Song et al. (2020) transformed metaphor identification, metaphor interpretation, and metaphor generation tasks into knowledge graph embedding tasks with triplets (source, attribute, target). The metaphor knowledge graph shared embeddings with concept-attribute collocations, which was only appropriate for nominal metaphors. Their used training data limited the vocabulary of generated concepts or attributes.

Su et al. (2021a) believed that the abstractness and concreteness of words were distinguishable in different modalities. They classified words as abstract and concrete concepts by concreteness values. They identified the metaphoricity of given word pairs for abstract concepts based on a logistic regression classifier and concreteness values. They additionally utilized image embeddings upon the concreteness values for concrete concepts to strengthen the concreteness representations in metaphor identification. However, the representations of abstract words were much more sparse than concrete ones in this work.

[21] https://osf.io/7emr6/.

[22] Supersenses are called "lexicographer classes" in WordNet documentation.

[23] A feed-forward neural network is a network where connections between nodes do not form a cycle, which differs from a recurrent neural network.

Ge et al. (2022) proposed a CMT-inspired multi-task learning framework based on the hypothesis that learning the source and target concept mappings could enhance the performance of metaphor identification. This work showed the state-of-the-art performance on word pair-based metaphor identification tasks, namely verb-noun and adjective-noun pairs. However, similar to other word pair-based methods, the CMT-inspired model needed to address the issue of identifying metaphoricity for MWEs. For example, "climb ladder" is a literal phrase without knowing other contextual information, while "*climb social ladder*" is a metaphor.

4.1.3 Token-level metaphor identification

Universal[24] **neural network.** Do Dinh and Gurevych (2016) used FNN on the metaphor detection task. They evaluated the performance of their model on the VUA dataset with different genre breakdowns, achieving positive results. Gao et al. (2018) proposed sequencing labeling and classification-based learning paradigms for token-level metaphor identification. Their experiments showed that standard bidirectional long short-term memory (BiLSTM) and additional contextual embeddings yielded strong performance for both learning paradigms on widely used metaphor identification datasets. These universal neural networks helped to reduce the reliance on features from background corpora, hand-coded rules, or additional manually created resources to some extent. However, these methods did not take advantage of the theoretical findings of metaphors.

External features. These approaches used novel features for processing the task. Since the use of metaphors changes among different levels of language proficiency, Stemle and Onysko (2018) trained the word embeddings with multiple learner corpora, simulating the human learning process. However, the learner data with low proficiency levels could result in linguistic errors and noises, such as grammar mistakes.

Gong et al. (2020) combined contextual information from the language models and multiple linguistic features from external resources, feeding into an FNN classifier for metaphor identification. The linguistic features included PoS, topic features (Klebanov et al. 2014), word concreteness (Brysbaert et al. 2014), WordNet features (Klebanov et al. 2016), VerbNet features (Klebanov et al. 2016), and corpus-based features (Klebanov et al. 2016). The ablation study showed that linguistic features could capture additional semantic information in addition to contextualized embeddings.

Kehat and Pustejovsky (2021) applied visibility embeddings to represent inputs. They hypothesized that physical language in multimodal datasets (texts and images) could be used for describing literal information because the physical language was comparatively more concrete. The visibility embeddings were constructed with words in different visual and non-visual corpora. Visibility embeddings were viewed as concreteness representations. However, their method captured representations for imitated concepts, which was insufficient for downstream metaphor identification tasks.

Ottolina et al. (2021) explored the influence of word embeddings in different periods on metaphor identification performance. They believed that the meanings of metaphors could change over time. However, the datasets that were used, namely TroFi (1985-1994) and VUA (1987-1989), have limited time spans compared with the examined embeddings (1900-2000).

Data augmentation. Some studies focused on solving the issues with inconsistency and sparse label annotation in datasets. Stowe et al. (2019) explored verb sense and training data

[24] "Universal" means the models are not metaphor-specific and can be applied in other linguistic tasks, such as sequence labeling or classification tasks.

from VerbNet (Schuler 2005) and learned syntactic patterns from Wikipedia. The additional data based on these patterns enhanced the original training dataset. However, the performance was only evaluated on verbs.

Previous methods failed to explicitly differentiate between a target word's metaphorical and literal senses. Lin et al. (2021) first combined semi-supervised learning with self-training to supplement metaphor datasets. Unlabelled data could be iteratively augmented in training data and generated pseudo-labels based on a contrastive learning objective to capture the distance between metaphorical and literal senses.

Yang et al. (2021) proposed two Sequence-to-Sequence (Seq2Seq) models (Long2Short and Short2Long) to generate additional training data based on syntactic patterns. They used dependency parsing tools to clean sentences, maintained specific PoS, and augmented data by Seq2Seq models. However, their method focused on verb metaphors only. The quality of the generated language based on their method was also unclear.

Linguistic theory basis. Linguistic metaphor-related theories, namely MIP and SPV, were essential in inspiring metaphor identification model design. Mao et al. (2019) constructed two models to learn metaphor identification based on MIP and SPV, respectively. Their models explicitly learned the semantic contrast between contextual and basic meanings (MIP) and the semantic contrast between a target word and its context (SPV), yielding state-of-the-art results. However, both methods modeled the semantic contrast in vector space rather than the natural language-based semantic contrast understanding in the original MIP and SPV theories.

Choi et al. (2021) leveraged MIP and SPV in one model. They employed two separate Transformer encoders (Vaswani et al. 2017) to acquire the representations of a target word in the sentence and the target word independent of the context. The MIP layer concatenated the vectors of the target word in the sentence with the target word alone for classification. The SPV layer concatenated the vectors of the target word in the sentence with the sentence for classification. The hidden states for two layers were concatenated together for the final prediction.

Qin and Zhao (2021) extracted frequently associated subjects and objects of the target words from Wikipedia corpus[25]. The vector representations of the original and the extracted subjects and objects were regarded as the contrast between contextual and basic meanings. These representations were learned in a Transformer-based neural network for classification.

Su et al. (2021b) utilized examples and definitions from the Oxford Dictionary[26] to make up the semantic information based on metaphor theories, namely SPV and MIP. The example model fine-tuned a pre-trained masked language model on the example sentences of a target word. The definition model simply concatenated multiple definitions of the target word with the original sentence, contrasting the literal and contextual meanings. However, their method did not distinguish the fact that the senses and example sentences of conventional metaphors were likely to be included in a dictionary. This could lower the model performance on conventional metaphors.

Chen et al. (2021) encoded sentences with and without metaphors in two ways after encoder layers to capture the semantically contextual inconsistency. For metaphorical sentences, they calculated the sum of the minimum distributional distances of target words against the other contextualized literal words as the inconsistency score. For literal sentences, they calculated the sum of the maximum distributional distances of every two words as the inconsistency score.

[25] http://nlp.stanford.edu/data/WestburyLab.wikicorp.201004.txt.bz2.

[26] https://www.lexico.com/.

Context relation. Many studies hypothesized that different context parts played different roles in identifying metaphors. For example, the position distance from contextual words to the target word determines the influence of contextual words on the metaphoricity of the target word to some extent. Su et al. (2020a) transformed the metaphor identification task into a reading comprehension paradigm. The whole sentence was considered as a global context. The sentence fragments were considered as a local context. The query word was considered as the question. Two Transformers processed the global and local context features separately but shared the same training weight parameters.

Zayed et al. (2020a) proposed a contextual modulation model on metaphor identification inspired by visual reasoning. It encoded a short phrase containing a target word and the whole sentence separately. An affine transformation processed their vector representations for the classification task.

Rohanian et al. (2020) first attempted metaphor identification with the awareness of MWEs. The model applied two graph convolutional neural networks (GCN) to integrate the dependency parsing information of the whole sentence and token-level representations of MWEs. MWEs were used as an extra input feature. The results showed that the MWE feature improved model performance on metaphor identification.

Song et al. (2021) extracted the grammatical local context, the sequential global context, and the basic meaning of a verb as a distant context hierarchically. They regarded a verb and its contexts as entities. Metaphor identification was regarded as relation classification between the verb and its corresponding context. This framework tried three vector combinations (concatenation, average, and maxout) and three relational modeling methods (linear, bilinear, and neural tensor models).

Li et al. (2021) considered neighboring sentences in metaphor identification and proposed a multi-level model. It hierarchically processed contextual information at the sentence level and discourse level. An early prediction of contextual labels was applied to facilitate the performance of this model. They assumed that contextual labels and other sentences in the dataset would also help to train the model. This model focused on contextual information, namely neighboring sentence representations and labels of contextual words, and might be less effective in short and single-sentence learning.

Interpretation assistance. Mao et al. (2018) combined metaphor identification and interpretation in an unsupervised learning fashion. Their model first predicted the best fit word most likely to appear in a context at a specific position via a Word2Vec-based language model. Secondly, if the cosine similarity between the best fit word and an original target word at the same position was lower than a specific threshold, the target word was identified as a metaphor in the context. This algorithm also reflected the idea of MIP that the contextual meaning of a metaphor is very different from its basic meaning, where the best fit word represented the contextual meaning. However, the dependency on WordNet made the word range and accuracy of WordNet limit the performance of this model.

Wan et al. (2021) proposed a multi-task learning framework for metaphor identification and interpretation based on word sense disambiguation. They extracted multiple glosses of the target words from the Merriam-Webster dictionary[27] and the Baidu Dictionary[28]. The glosses were encoded with the input sentences for metaphor identification. The representations after the gloss encoder combined with input sentence vectors were encoded by attention for interpretation. However, the annotation resource was limited in this work.

[27] https://www.merriam-webster.com/dictionary/.

[28] https://dict.baidu.com.

Multi-task learning. Chen et al. (2020) proposed a multi-task learning model, which identified metaphors and idioms simultaneously. They argued that the representations of out-of-domain data and the similarity of different figurative languages could enhance the performance on metaphor identification.

Le et al. (2020) studied word sense disambiguation and metaphor identification in a multi-task learning model. They believed that both tasks were related to human cognition. The model contained a GCN to organize context dependency relations. It removed context that did not have a direct dependency relation with the target word, which reduced the training time.

Mao and Li (2021) studied a metaphor identification (main) task together with a PoS tagging (auxiliary) task because a PoS label was a practical feature for metaphor identification. They proposed a novel soft-parameter sharing mechanism, the Gated Bridging Mechanism, to enhance the learning of the main task. The motivation was that Gated Bridging Mechanism could filter out useless information and receive supportive information from an auxiliary task. Their model has been embedded in MetaPro (Mao et al. 2022a) as the metaphor identification module, achieving state-of-the-art results in both open-class word-based and all-PoS-based metaphor identification tasks.

4.1.4 Summary

Table 6 shows the task definitions and PoS of recent metaphor identification works. Most studies focused on verb metaphors, as verb metaphors occur more frequently than others in texts. The relation-level metaphor identification studies also examined adjective metaphors and noun metaphors since adjective-noun metaphor pairs are a component of metaphor families (Krishnakumaran and Zhu 2007). The token-level SEQ studies identified metaphors in all PoS. The output could be more conveniently employed in downstream tasks. Thus, the token-level SEQ-based methods have recently attracted the most attention in the community. Only a few studies focused on multi-lingual metaphor identification, indicating a limited transformation of current models into other languages.

Table 7 shows the features employed in metaphor identification models. Here, we use "Word emb." to denote static word embedding methods independent of contextual information in downstream tasks, such as Word2Vec and GloVe. A pre-trained language model (PLM) denotes a context-dependent word embedding method, such as ELMo (Peters et al. 2018), BERT (Kenton and Toutanova 2019), and RoBERTa (Liu et al. 2019). The top three most popular features used in metaphor identification were POS, PLM, and word embedding. As PLM shows excellent sentence representation skills, many researchers in recent years have chosen PLM to replace word embedding. Compared with the other three categories, the token-level CLS studies (Do Dinh and Gurevych 2016; Su et al. 2020a; Le et al. 2020) used explicit position information as features because CLS-based methods predicted the label of a target word in a sentence. The position information indicates the position of the target word in a sentence. Semantic attributes, such as concreteness, imageability, and affect scores, were more prevalent in sentence-level and relation-level models. These features have linguistic and cognitive intuition. However, the token-level methods began to be abandoned due to the growth of pre-trained language models that contain richer contextual information. Lexical resources were also less attractive in token-level methods for the same reason. External knowledge, such as commonsense and semantic knowledge, delivers information beyond context. Thus, several token-level studies (Mao et al. 2018; Stowe et al. 2019; Gong et al. 2020) still used features from lexical resources in their models. Visual features were also employed in metaphor identification tasks (Shutova et al. 2016; Su et al. 2021a; Kehat and Pustejovsky 2021) because they contain additional commonsense knowledge.

Table 6 Metaphor identification models by author, task definition, and PoS

Category	Paper	Verb	Adj.	Noun	Adv.	All	Multi -lingual
Sentence -level	Krishnakumaran and Zhu (2007)	✓	✓	✓			
	Tsvetkov et al. (2013)	✓					✓
	Mohler et al. (2013)					✓	
	Wan et al. (2020)	✓					
Relation -level	Tsvetkov et al. (2014)	✓	✓				✓
	Shutova et al. (2016)	✓	✓	✓			
	Bulat et al. (2017)		✓				
	Rei et al. (2017)	✓	✓				
	Song et al. (2020)		✓	✓			✓
	Su et al. (2021a)	✓	✓	✓			
	Ge et al. (2022)	✓	✓				
Token -SEQ	Stemle and Onysko (2018)					✓	
	Gao et al. (2018)-SEQ	✓				✓	
	Mao et al. (2019)	✓				✓	
	Stowe et al. (2019)-SEQ	✓				✓	
	Gong et al. (2020)	✓				✓	
	Mao and Li (2021)(MetaPro)					✓	
	Chen et al. (2021)	✓				✓	✓
	Kehat and Pustejovsky (2021)	✓				✓	
	Li et al. (2021)	✓				✓	
	Ottolina et al. (2021)	✓				✓	

Table 6 continued

Category	Paper	Verb	Adj.	Noun	Adv.	All	Multi -lingual
Token -CLS	Do Dinh and Gurevych (2016)	✓	✓	✓	✓		
	Mao et al. (2018)	✓					
	Gao et al. (2018)-CLS	✓					
	Stowe et al. (2019)-CLS	✓					
	Su et al. (2020a)	✓				✓	
	Zayed et al. (2020a)	✓	✓				
	Chen et al. (2020)	✓				✓	
	Le et al. (2020)	✓				✓	
	Rohanian et al. (2020)	✓					
	Lin et al. (2021)	✓				✓	
	Yang et al. (2021)	✓					✓
	Song et al. (2021)	✓	✓	✓	✓	✓	
	Choi et al. (2021)	✓				✓	
	Wan et al. (2021)	✓				✓	✓
	Su et al. (2021b)	✓					
	Qin and Zhao (2021)	✓					

Adj. adjective, *Adv.* adverb

Table 7 Metaphor identification models by author and feature

Category	Paper	POS	Posit.	Cooc.	Concr.	Sem. attr.	Lex. Res.	Word emb.	PLM	Graph emb.	Visual
Sentence -level	Krishnakumaran and Zhu (2007)	✓		✓			✓				
	Tsvetkov et al. (2013)	✓			✓	✓	✓				
	Mohler et al. (2013)			✓			✓				
	Wan et al. (2020)	✓				✓		✓			
Relation -level	Tsvetkov et al. (2014)	✓			✓	✓	✓	✓			
	Shutova et al. (2016)	✓						✓			✓
	Bulat et al. (2017)	✓				✓		✓			
	Rei et al. (2017)	✓						✓			
	Song et al. (2020)	✓					✓			✓	
	Su et al. (2021a)	✓			✓			✓			✓
	Ge et al. (2022)	✓		✓			✓		✓		
Token -SEQ	Stemle and Onysko (2018)	✓						✓			
	Gao et al. (2018)-SEQ							✓	✓		
	Mao et al. (2019)							✓	✓		
	Stowe et al. (2019)-SEQ	✓					✓	✓	✓		
	Gong et al. (2020)	✓			✓	✓	✓		✓		
	Mao and Li (2021)(MetaPro)								✓		
	Chen et al. (2021)							✓	✓		
	Kehat and Pustejovsky (2021)								✓		✓
	Li et al. (2021)		✓	✓					✓		
	Ottolina et al. (2021)							✓			

Table 7 continued

Category	Paper	POS	Posit.	Cooc.	Concr.	Sem. attr.	Lex. Res.	Word emb.	PLM	Graph emb.	Visual
Token -CLS	Do Dinh and Gurevych (2016)	✓			✓			✓			
	Mao et al. (2018)	✓	✓				✓	✓			
	Gao et al. (2018)-CLS	✓						✓	✓		
	Stowe et al. (2019)-CLS	✓					✓	✓	✓		
	Su et al. (2020a)	✓	✓						✓		
	Zayed et al. (2020a)	✓						✓	✓		
	Chen et al. (2020)								✓		
	Le et al. (2020)	✓						✓	✓		
	Rohanian et al. (2020)	✓							✓		
	Lin et al. (2021)			✓					✓		
	Yang et al. (2021)	✓									
	Song et al. (2021)	✓							✓		
	Choi et al. (2021)		✓						✓		
	Wan et al. (2021)						✓		✓		
	Su et al. (2021b)	✓					✓		✓		
	Qin and Zhao (2021)	✓		✓				✓			

Posit. denotes position, *Cooc.* co-occurrence, *Concr.* concreteness, *Sem. attr.* semantic attribute, *Lex. Res.* lexical resource, *emb.* embedding, *PLM* pre-trained language model

Table 8 shows the frameworks employed in metaphor identification studies. Metaphor identification, as a relatively active task in NLP, has constantly been applying the most up-to-date computational techniques in these years. Most sentence-level and relation-level studies (Tsvetkov et al. 2013; Mohler et al. 2013; Bulat et al. 2017) utilized traditional machine learning classifiers, such as logistic regression, random forest, and support vector machine. Later, more studies (Stemle and Onysko 2018; Gao et al. 2018; Mao et al. 2019) introduced LSTM, Attention, and Transformer into their models for contextualized learning. Graph learning (Song et al. 2020; Le et al. 2020; Rohanian et al. 2020) and multi-task learning (MTL) (Chen et al. 2020; Le et al. 2020; Wan et al. 2021; Mao and Li 2021; Ge et al. 2022) have started to appear in metaphor identification because these paradigms can fuse additional information with different data structures, relevant tasks, and information sharing methods.

Table 9 demonstrates the evaluation setups and results on the primary datasets. The numbers are F1 scores in the test set. Studies on relation-level metaphor identification used word pair datasets, such as MOH and TSV. Ge et al. (2022) obtained the best performance on relation-level metaphor identification tasks, namely 75.6% F1 in MOH and 86.6% F1 in TSV. The most popular datasets in token-level tasks are MOH-X, TroFi, VUA-V, and VUA-A. Lin et al. (2021) achieved the highest F1 score in MOH-X, 84.7%. The top result in TroFi was 89.3%, achieved by Wan et al. (2021). Yang et al. (2021) achieved the highest performance on VUA-V, reaching 80.65%. Choi et al. (2021) outperformed others in VUA-4 with an F1 score of 79.8%. Mao et al. (2022a) yielded the highest F1 score in an all-PoS metaphor identification task (VUA-A), reaching 79.2% F1 scores.

By cross-referencing multiple tables above, we can observe that all the best performers (Lin et al. 2021; Yang et al. 2021; Choi et al. 2021; Wan et al. 2021; Mao et al. 2022a) in token-level tasks utilized PLMs. Based on their learning paradigms, these models also processed POS, word co-occurrence, or lexical resources as supplements. Mao et al. (2022a) applied an MTL framework, jointly learning metaphor identification and PoS tagging tasks, and achieved the best performance on VUA-A. Previous works demonstrated that incorporating PoS tags, whether presented as explicit features or acquired through MTL, could improve the detection of metaphors.

Metaphor identification has attracted much attention due to two shared tasks (Leong et al. 2018, 2020) and a large-scale annotated corpus (Steen et al. 2010a). However, identifying other types of metaphors, such as extended metaphors or MWEs, has yet to be well solved. Extended metaphors include multiple source and target concept mappings with close associations. Identifying extended metaphors requires a perspective of the whole text and connections among different concept mappings. Metaphorical MWEs are inseparable in the text, which is challenging in current token-level metaphor identification models.

4.2 Metaphor interpretation

Current metaphor interpretation tasks can be categorized into three types, namely property extraction, word-level paraphrasing, and explanation pairing.

(1) Property extraction-based metaphor interpretation systems aim to extract properties that can link source and target domains. The properties represent the shared features of two words from source and target domains, e.g., extracting an adjective property for two nouns from source and target domains.

(2) Word-level paraphrasing systems aim to paraphrase a metaphor into its literal counterpart within the context word by word. Typically, these systems can paraphrase single-word metaphors.

Table 8 Metaphor identification models by author and learning paradigm

Cate- gory	Paper	Rule	Cos.	Unsup.	Semi -sup.	LR	DT	RF	SVM	RM	FNN	LSTM	Attn.	XFMR	Gat.	Graph	MTL	Data	PTM	S2S
Sent. -level	Krishnakumaran and Zhu (2007)	✓																		
	Tsvetkov et al. (2013)					✓														
	Mohler et al. (2013)						✓		✓											
	Wan et al. (2020)											✓	✓							
Rel. -level	Tsvetkov et al. (2014)							✓												
	Shutova et al. (2016)		✓																	
	Bulat et al. (2017)								✓											
	Rei et al. (2017)		✓												✓					
	Song et al. (2020)															✓				
	Su et al. (2021a)		✓			✓														
	Ge et al. (2022)										✓						✓			
Token -SEQ	Stemle and Onysko (2018)											✓								
	Gao et al. (2018)-SEQ											✓								
	Mao et al. (2019)											✓	✓							
	Stowe et al. (2019)-SEQ											✓						✓		
	Gong et al. (2020)										✓									
	Mao and Li (2021)(MetaPro)													✓	✓		✓			
	Chen et al. (2021)											✓								
	Kehat and Pustejovsky (2021)											✓								
	Li et al. (2021)												✓	✓						
	Ottolina et al. (2021)											✓								

Table 8 continued

Cate- gory	Paper	Rule	Cos.	Unsup.	Semi -sup.	LR	DT	RF	SVM	RM	FNN	LSTM	Attn.	XFMR	Gat.	Graph	MTL	Data	PTM	S2S
Token -CLS	Do Dinh and Gurevych (2016)										✓									
	Mao et al. (2018)		✓	✓																
	Gao et al. (2018)-CLS											✓								
	Stowe et al. (2019)-CLS											✓						✓		
	Su et al. (2020a)													✓						
	Zayed et al. (2020a)										✓	✓	✓							
	Chen et al. (2020)										✓						✓			
	Le et al. (2020)											✓				✓	✓			
	Rohanian et al. (2020)												✓			✓				
	Lin et al. (2021)				✓													✓		
	Yang et al. (2021)																	✓		✓
	Song et al. (2021)									✓										
	Choi et al. (2021)													✓						
	Wan et al. (2021)										✓						✓			
	Su et al. (2021b)										✓								✓	
	Qin and Zhao (2021)										✓		✓	✓						

“Rule” denotes a rule-based method

Cos. cosine similarity, *Unsup.* unsupervised learning, *Semi-sup.* semi-supervised learning, *LR* logistic regression, *DT* decision tree, *RF* random forest, *SVM* support vector machine, *RM* a relation model, *Attn.* attention, *XFMR* Transformer, *Gat.* a gati0ng mechanism, *Graph* a graph-based method, *MTL* multi-task learning, *Data* a data-driven method. *PTM* denotes a pre-trained model, *S2S* Seq-to-Seq

Table 9 Metaphor identification models by author and evaluation

Cate- gory	Paper	Self -collected	MOH	MOH -X	TSV	TroFi -SVO	TroFi	VUA -V	VUA -4	VUA -A	TOEFI -V	TOFEL -A	CMRSA	Other
Sent. -level	Krishnakumaran and Zhu (2007)													MML
	Tsvetkov et al. (2013)						78.0%							
	Mohler et al. (2013)	✓												
	Wan et al. (2020)							76.1%						
Rel. -level	Tsvetkov et al. (2014)				85.0%	79.0%								
	Shutova et al. (2016)		75.0%		79.0%									
	Bulat et al. (2017)				77.0%									
	Rei et al. (2017)		74.2%		81.1%									
	Song et al. (2020)	✓												
	Su et al. (2021a)				85.0%	**83.0%**								MOH-VN-68%, GUT-85.0%
	Ge et al. (2022)		**75.6%**		**86.6%**									
Token -SEQ	Stemle and Onysko (2018)								61.3%					
	Gao et al. (2018)-SEQ			75.6%			71.1%			72.6%				
	Mao et al. (2019)			80.0%			72.4%	70.8%		74.3%				
	Stowe et al. (2019)-SEQ			70.4%			68.4%	69.6%		73.8%				
	Gong et al. (2020)							77.1%	73.0%		71.9%	70.3%		
	Mao and Li (2021)									76.9%				
	Chen et al. (2021)			80.5%			73.3%	71.7%	74.8%				69.5%	
	Kehat and Pustejovsky (2021)			84.6%				73.6%		77.4%				
	Li et al. (2021)							76.8%		73.0%				
	Ottolina et al. (2021)			81.0%			74.0%			74.0%				
	Mao et al. (2022a)								73.1%	**79.2%**				

Table 9 continued

Cate- gory	Paper	Self -collected	MOH	MOH -X	TSV	TroFi -SVO	TroFi	VUA -V	VUA -4	VUA -A	TOEFI -V	TOFEL -A	CMRSA	Other
Token -CLS	Do Dinh and Gurevych (2016)								57.5%					
	Mao et al. (2018)		74.0%	75.0%										
	Gao et al. (2018)-CLS			79.1%			72.0%			58.9%				
	Stowe et al. (2019)-CLS			70.4%			68.4%	67.9%						
	Su et al. (2020a)			84.2%			76.1%	80.4%	76.9%		**74.9%**	**71.5%**		
	Zayed et al. (2020a)			77.9%	83.4%		87.9%	68.3%						ZayTw-78.6%
	Chen et al. (2020)							77.5%	73.4%		70.2%	69.2%		
	Le et al. (2020)			79.6%			73.2%	71.7%		75.1%				
	Rohanian et al. (2020)			80.2%			72.8%							
	Lin et al. (2021)			**84.7%**			74.5%	75.6%		79.0%				
	Yang et al. (2021)							**80.7%**					**86.4%**	
	Song et al. (2021)			84.2%			72.9%	75.9%	70.9%	77.2%				
	Choi et al. (2021)			79.2%			62.0%	77.1%	**79.8%**					
	Wan et al. (2021)						**89.3%**	75.0%		77.2%				PSUCMC-V-71.2, PSUCMC-A-78.7
	Su et al. (2021b)			83.4%			71.4%	76.1%						
	Qin and Zhao (2021)			81.6%			74.3%							

The best performance (F1 score) in each column is in bold. VUA-V verb metaphors in VUA, *VUA-4* verb, adjective, noun, and adverb metaphors (open-class words) in VUA, *VUA-A* all PoS in VUA. *TroFi-SVO* subject-verb and verb-object pairs, *CMRSA* Chinese Metaphor Recognition and Sentiment Analysis, *MOH-VN* verb-noun pairs in MOH, *PSUCMC* the PSU Chinese Metaphor Corpus

Table 10 Examples of metaphor interpretation forms

Category	Metaphor	Interpretation
Property extraction (Su et al. 2017)	Love is *tide*	Love is unstoppable
Word-level paraphrasing (Mao et al. 2018)	She *devoured* his novels.	She enjoyed his novels
Explanation pairing (Mao et al. 2022a)	This is *a red letter day*.	This is a red letter day, where “a red letter day” means a day of significance.

(3) Explanation pairing-based systems aim to pair an explanation with a metaphor in a context. Usually, the explanations come from dictionary definitions.

Examples of the metaphor interpretation tasks are shown in Table 10.

4.2.1 Property extraction

Su et al. (2015) proposed a model extracting the properties of source and target domains in a Chinese nominal metaphor interpretation task. They selected the candidate properties of source concepts from Attribute Database[29] and Sardonicus[30], and extended them with synonyms from Tongyi Cilin (Extended)[31]. They defined semantic relatedness by combining the cosine similarity between the properties and target concepts with the cosine similarity between the properties and context words. The output interpretation was designated as “target be property”, where the property word achieved the highest relatedness score. However, this method failed to process sentences with complex syntactic structures. Instead of calculating the cosine similarity between the candidate properties and target concepts, Su et al. (2017) revised the relatedness score mentioned above into an average of cosine similarities between the synonyms of the properties and target concepts. The context words played a minor part in this model compared to the work of Su et al. (2015).

Rai et al. (2019) proposed an unsupervised metaphor interpretation method with emotion analysis. This work employed emotion as a connection between source and target concepts. The hypotheses were that metaphors could be better understood from the perspective of emotion; metaphors had multiple senses in different perceptions. The model extracted related properties of source concepts from the web and vectorized each property in six emotion dimensions: anger, fear, happiness, disgust, sadness, and surprise. The algorithm factored the closeness of properties and the target concept, and the emotion tendency of properties on one of six emotion dimensions. The final output could be briefly expressed as “target be <emotion: property>.”

Su et al. (2020b) proposed a culture-related hierarchical semantic model for interpreting Chinese nominal metaphors. They hypothesized that some metaphors were generated with cultural connotations. To extend the property scope, they manually annotated a concept mapping knowledge base for culture-related concepts. They trained culture-semantic vectors with famous Chinese literary works via Word2Vec. They utilized a random walk algorithm to obtain the most appropriate attributes with the highest relevance to the target concepts.

[29] A database by NLP Lab of Xiamen University.

[30] An adjective taxonomy database from https://afflatus.ucd.ie/.

[31] A Chinese Thesaurus, http://ir.hit.edu.cn/.

Song et al. (2020) transformed metaphor interpretation into a knowledge graph completion task. The knowledge graph was built with (source, attribute, target) triplets. The attribute represented a shared property between source and target concepts. The task was selecting an attribute based on the source and target concepts. The final attributes were extracted from a set of candidates. The evaluation included exact-match-based and synonymy-based metrics to show a comprehensive model performance.

To sum up, the methods above only worked for nominal metaphor interpretation tasks. Interpreting properties could simplify the relationship between source and target nouns, whereas the property-based methods could hardly expand to other PoS of metaphors, such as verbs, adjectives, and adverbs. The interpretation outputs were formulated as “target be property”, which was not convenient for supporting downstream tasks, given the complexity of syntax in real-world text.

4.2.2 Word-level paraphrasing

Shutova (2010) first defined metaphor interpretation as a paraphrasing task for word pairs: verb-direct object and verb-subject. The method first selected possible candidates from a large corpus by the co-occurrence of targets and specific syntactic dependency relations. Next, it filtered and ranked them by hyponyms in WordNet and selectional preference.

Mao et al. (2018) built an unsupervised method for word-level paraphrasing from whole sentences to improve usability and support downstream tasks and language learners. They hypothesized that the literal senses of words occurred more frequently than their metaphorical senses in corpora. Thus, they utilized synonyms and hypernyms from WordNet as the candidate literal senses of the target words. They also found that Word2Vec input and output vectors could better represent words and context co-occurrences. This model was evaluated as a pre-processing technique for machine translation tasks, demonstrating that machine translation systems such as Google and Bing translators could be improved mainly by this model. Later, Mao et al. (2022a) extended this method in MetaPro using a pre-trained language model instead of Word2Vec. Mao et al. (2022a) demonstrated that metaphor processing could improve state-of-the-art sentiment analysis classifiers. However, due to the model’s reliance on WordNet, its performance has been restricted by the word range and accuracy of WordNet.

Word-level paraphrasing-based methods can be used as a text pre-processing technique, improving the semantic understanding for diverse downstream NLP tasks and second language acquisition. However, current word-level paraphrasing methods failed to capture the nuance between metaphors and their literal counterparts. Furthermore, many metaphors are MWEs, such as idioms, where word-level paraphrasing methods could not handle these cases well.

4.2.3 Explanation pairing

Martin (1990) proposed a metaphor interpretation, denotation, and acquisition system (MIDAS). The interpretation system first parsed the input sentence and obtained the syntactic information. Possible candidates were collected based on syntax and validated with coherence and abstractness in constraint checking. After a recursive procedure, the output was an interpretation as an explanation of the metaphorical expression. This system processed metaphors with grammar rules in language and rich knowledge.

Bizzoni and Lappin (2018) proposed a CNN-LSTM framework to capture the semantic representations of metaphor and literal expressions. The input of the model was a metaphorical sentence and a literal candidate. The top-ranked candidate was the paraphrase of a metaphorical sentence. This model needed a large size of tailored annotated data, namely candidate paraphrases, which was not convenient in real-world applications.

Mao et al. (2022a) additionally proposed a dictionary and rule-based method to identify and interpret metaphorical MWEs from sentences to mitigate the limitations of word-level paraphrasing-based methods. They used dependency triplets (a head word, dependency, a tail word) and lemmas as features, pairing MWEs in a source sentence with MWEs in their predefined feature-mapping dictionaries. They also defined a dictionary that contained multiple explanations for every MWE. Next, given a paired MWE (an identified metaphorical MWE), the selected explanation was given by comparing the semantic similarity between an explanation and the source sentence. Finally, they concatenated the explanation of an MWE with the source sentence as a clause (see Table 10). Thus, downstream tasks could directly use the output in a pre-processing fashion. This method outperformed machine learning baselines in an idiomatic MWE detection task. The metaphorical MWE interpretation also brought about improvement in sentiment analysis tasks. However, this model depended on collected knowledge for limited metaphors, which made it hard to update and include a more extensive range of metaphorical MWEs.

4.2.4 Summary

Table 11 shows the task definition and research scopes in metaphor interpretation. Noun metaphors obtained more attention than verb metaphors in relation-level metaphor interpretation tasks because these studies largely depended on verb-noun and adjective-noun metaphor datasets. Verb metaphor interpretation was usually studied at word-level paraphrasing because another literal counterpart verb could relatively intuitively paraphrase a metaphorical verb. However, there was a nuance between the original metaphor and the paraphrase. Property extraction was a frequent manifestation of noun metaphor interpretation, as we have seen that the properties of source and target nouns can be represented as adjectives. The work of Mao et al. (2022a) can interpret metaphors in verbs, nouns, adjectives, adverbs, and MWEs, significantly improving the functionality and usability of a metaphor interpretation system in different application scenarios.

Table 12 illustrates the features used in interpretation studies. We find that many studies relied on lexical resources. It highlights that the analogical ability of pure machine intelligence is still weak. Metaphor interpretation is a high-level linguistic understanding and reasoning task. Thus, lexical resources were employed as an additional feature to provide knowledge beyond input texts to achieve accurate metaphor interpretation. We also find that many studies examined the task by PoS, either using a metaphor dataset with a specific PoS or directly using PoS as input features. Such a pattern shows the diversity of approaches to interpreting metaphors because there are different interpretation methods for metaphors with different PoS. Many word-level studies interpreted metaphors by learning the association of words and contexts. They either used the co-occurrence statistics (Shutova 2010), Word2Vec input and output vectors (Mao et al. 2018), or a PLM-based masked word prediction method (Mao et al. 2022a). Exploiting concept mapping (Su et al. 2020b) proved that conceptual information could assist interpretation tasks.

Table 13 shows the learning paradigms in metaphor interpretation studies. Given insufficient annotated metaphor interpretation datasets, multiple studies delivered the task in an

Table 11 Metaphor interpretation models by author, task definition, and research scope

Category	Paper	Rel.-level	Sent.-level	Verb	Noun	Adj.	Adv.	MWE	SD
Property extraction	Su et al. (2015)		✓		✓				
	Su et al. (2017)	✓			✓				
	Rai et al. (2019)	✓			✓				
	Su et al. (2020b)		✓		✓				✓
	Song et al. (2020)	✓			✓				
Word-level paraphrasing	Shutova (2010)	✓		✓					✓
	Mao et al. (2018)		✓	✓					
	Mao et al. (2022a)		✓	✓	✓	✓	✓		
Explanation pairing	Martin (1990)		✓	✓					
	Bizzoni and Lappin (2018)		✓					✓	
	Mao et al. (2022a)		✓					✓	

SD specific domain

Table 12 Metaphor interpretation models by author and features

Category	Paper	POS	Cooccu.	Lex. Res.	Word emb.	Graph emb.	Concept mapping	PLM
Property extraction	Su et al. (2015)	✓		✓	✓			
	Su et al. (2017)	✓		✓	✓			
	Rai et al. (2019)	✓		✓				
	Su et al. (2020b)	✓		✓	✓		✓	
	Song et al. (2020)	✓		✓		✓		
Word-level paraphrasing	Shutova (2010)	✓	✓	✓				
	Mao et al. (2018)	✓		✓	✓			
	Mao et al. (2022a)	✓		✓				✓
Explanation pairing	Martin (1990)			✓				
	Bizzoni and Lappin (2018)				✓			
	Mao et al. (2022a)	✓		✓				

unsupervised learning manner. Apart from deep learning methods (Mao et al. 2022a), graph-based (Song et al. 2020; Su et al. 2020b) and rule-based (Martin 1990; Mao et al. 2022a) methods are still active in the community. Graph-based methods are practical tools to search for appropriate properties among candidates in addition to linguistic similarity-based methods (Su et al. 2015, 2017).

Table 14 illustrates the evaluation settings of current metaphor interpretation studies. Due to the lack of gold metrics to measure the interpretation task, automatic evaluation metrics such as accuracy and mean reciprocal rank were employed in word-level paraphrasing and explanation pairing tasks. Most studies also used human evaluation (Rai et al. 2019; Su et al. 2020b; Mao et al. 2022a) to assess the interpretation performance. These authors asked participants to evaluate the metaphor interpretation outputs by acceptability, appropriateness, and metaphoricity. For example, Mao et al. (2022a) believed that coherence, semantic completeness, and literality were three critical evaluation dimensions for a successful paraphrasing-based system.

5 Conceptual metaphor processing

Current conceptual metaphor processing methods target mapping the source and target concept domains of metaphors. Commonly, concept mappings are represented in the form of “a target concept is a source concept,” such as “ARGUMENT IS WAR.” Conceptual metaphor processing methods can be categorized into three types:

(1) Clustering-driven methods define the source and target domains as word clusters, where the concepts are manually named according to the words and entities in the same clusters.
(2) Conceptualization-driven methods obtain fine-grained concept mappings based on syntactic patterns and lexical information. A fine-grained concept agent can only represent narrow entities and limited words by semantic relations. Usually, it is obtained according to syntactic dependency constraints and semantic coherence to the context.
(3) Abstraction-driven methods automatically abstract concept agents from identified source and target entities. Compared with conceptualization-driven methods, whose concepts are represented by semantically related words, abstraction-driven methods tend to generate abstract concept agents representing a group of fine-grained concepts, where conceptual relations connect the fine-grained concepts.

A clear difference between conceptualization-driven and abstraction-driven methods is that concepts generated from abstraction-driven methods are more abstract and representative than those from conceptualization-driven methods. In Example 2.8, Li et al. (2013) argued that a fine-grained concept mapping for “car” is “horse” because “horse” is semantically and syntactically coherent to the context verb “drink” in a subject-verb dependent relationship. However, “horse” is a particular entity, which cannot represent conceptual relations between a subject and the verb “drink”. Thus, abstraction-driven methods aim to further abstract concept agents from source and target entities to represent the concept mappings, such as “animal” for “car” in Example 2.8. In other words, conceptualization-driven methods use source and target entities as concept agents. In contrast, abstraction-driven methods take one step further, abstracting general concepts as agents from the source and target entities.

Table 13 Metaphor interpretation models by author and learning paradigm

Category	Paper	Cos.	Unsup.	DL	Graph -based	Rule -based
Property extraction	Su et al. (2015)	✓	✓			
	Su et al. (2017)	✓	✓			
	Rai et al. (2019)	✓	✓			
	Su et al. (2020b)				✓	
	Song et al. (2020)				✓	
Word-level paraphrasing	Shutova (2010)		✓			
	Mao et al. (2018)	✓	✓			
	Mao et al. (2022a)	✓	✓	✓		
Explanation pairing	Martin (1990)					✓
	Bizzoni and Lappin (2018)			✓		
	Mao et al. (2022a)	✓				✓

DL deep learning models

Table 14 Metaphor interpretation models by author and evaluation

Category	Paper	Auto.	HMN	Acc.	Accept.	MRR	Hits@N	Others
Prop. extra.	Su et al. (2015)		✓		✓			
	Su et al. (2017)		✓		✓			
	Rai et al. (2019)		✓					APROP
	Su et al. (2020b)		✓		✓			
	Song et al. (2020)	✓				✓	✓	MR
Word-l. paraph.	Shutova (2010)	✓		✓		✓		
	Mao et al. (2018)		✓		✓			Coh
	Mao et al. (2022a)		✓					Coh, SC, Lit
Explan. pairing	Martin (1990)							
	Bizzoni and Lappin (2018)	✓		✓				Corr.
	Mao et al. (2022a)	✓		✓				

Auto. automatic evaluation. *HMN* human evaluation, *Acc.* accuracy, *Accept.* acceptability, *MRR* Mean Reciprocal Rank, *APROP* appropriateness, *MR* mean rank, *Corr.* correlation. *Coh* coherence. *SC* semantic completeness, *Lit* literality

5.1 Clustering-driven methods

Mason (2004) proposed a system called CorMet to find concept mappings for conventional metaphors. It started with two specific domains, searched documents related to the domains, and subsequently extracted representative verbs and clustered selectional preferences in WordNet nodes. The step of searching documents online can introduce unnecessary noise for representative verbs. Those verbs might not be related to the specific domains. The test was completed on a subset of MML due to the lack of large corpora with golden labels.

Gandy et al. (2013) used a rule-based heuristic algorithm to generate concept mappings. They first utilized lexical information, such as frequent collocations, abstractness scores, and semantic categories from WordNet to identify linguistic metaphors. They extracted each target's metaphor candidates with positive point-wise mutual information (PMI). They measured the similarity between each candidate's facets (expressing aspects of source domains) and aimed to discover nominal analogies. They finally clustered candidates and self-defined an overall score with abstractness and facet distribution to select a meaningful concept from the WordNet hypernyms. This model attempted to use as few domain-specific or language-specific knowledge bases as possible for a straightforward generalization. However, the method was only evaluated on "God", "governance", "government", "mother", and "father" domains.

Strzalkowski et al. (2013) utilized topical structures, imageability scores, and a clustering method to generate concept mappings from a passage in a specific domain. They hypothesized that metaphorical words were usually used outside the topical structure of a sentence and with high imageability. They also introduced affect and force (commonness) to measure the different aspects of a metaphor.

Clustering-driven methods could categorize words into clusters by semantic meanings. The clusters represented groups of words, whereas the groups of words were not abstracted or conceptualized. In other words, a concept could not automatically represent a group. As a result, the groups of words and the corresponding mappings given by clustering-driving methods could not be interpreted without manual efforts. Besides, previous works mainly depended on domain-specific knowledge to develop the clusters, leading to difficulties in generalization.

5.2 Conceptualization-driven methods

Li et al. (2013) proposed the first big data-driven and unsupervised metaphor detection and concept generation method. They obtained metaphorical and literal word pair corpora from a web corpus with like-a and is-a syntactic patterns. Each pair in their corpora contained frequency-based probability scores. They designed a context-based formula to determine the implicit source or target concepts among the candidates from the metaphorical corpus. However, the simple learned dependency limited its applications to expressions with diverse syntactic structures.

Gagliano et al. (2016) applied Word2Vec to find the connection between two words in the semantic space to generate a figurative relationship. The addition and intersection models obtained candidates of a source concept according to Word2Vec vectors of a target concept and an attribute. Workers from AMT selected the best choice of source concept. The analysis showed that the candidate concepts whose cosine similarity between two words was balanced could enhance the blending effect of two semantic spaces.

Rosen (2018) proposed a source domain mapping model and extracted features based on dependency relationships to express the interaction between a target word and its construction-based context. The source domains were performed as one-hot vectors in neural networks, which may cause data sparsity. The model could only output one of 77 available and limited source domains.

Conceptualization-driven methods could generate fine-grained mappings for metaphors. The conceptualization processes aimed to seek an appropriate source or target concept. The concept mappings could only represent limited metaphors in these methods, so they showed restricted roles in downstream tasks.

5.3 Abstraction-driven methods

Dodge et al. (2015) proposed a system to bridge theory-driven and corpus-driven methods of metaphor identification. They prepared a hand-crafted repository with linguists. The repository contained concept domains and multiple relations such as “subcase of” and “incorporates as a role”. Subsequently, they searched sentences by a set of grammatical constructions. The source and target words were matched with concept domains in the repository via its relational network with the help of WordNet, FrameNet[32], and Wiktionary[33]. This work was limited to detecting metaphors by pre-defined syntactic patterns from linguistic features and knowledge bases.

Fu et al. (2020) defined an image concept mapping task with three components: literal concept detection, literal-implied concept mapping, and metaphor captioning. The first part was to detect textual concepts from images. The second aimed to find suitable target concepts based on source and contextual concepts from the first part. The third part was to generate a caption based on all concepts. This work focused on the second part and assumed that the concepts were readily detected. Though the experiment dataset contained images, the proposed model and chosen baselines all processed text annotated from images. As a result, we consider this work as textual metaphor processing research. They built an undirected reference graph, where nodes were candidates obtained by searching with rule-based queries, and edges reflected the compatibility between candidates and contextual concepts. The final implied concepts had the highest compatibility among all the concepts.

Ge et al. (2022) proposed a multi-task learning model with a dynamic reward mechanism for metaphor identification and concept mapping generation. They proposed a WordNet and knee algorithm (Satopaa et al. 2011)-based method for abstracting concepts from words. Due to the absence of large annotated concept mapping datasets, the dynamic reward mechanism mitigated this issue by pushing the learned concept mappings toward more accurate metaphor identification. Because there were richer resources for supervised learning metaphor identification, the model could learn concept mappings with a broad conceptual spectrum from metaphor identification. Compared with previous works, the advantage of this model was that the model could automatically abstract concepts and generate concept mappings in natural language. However, this framework for concept mappings failed to be adjusted by different contexts and more complicated syntactic structures.

Mao et al. (2023)[34] integrated the work of Ge et al. (2022) to process concept mapping from end to end. The system first identified metaphors from an input sentence. Next, the identified metaphors were paraphrased into their literal counterparts. The source and target

[32] https://framenet.icsi.berkeley.edu.

[33] https://www.wiktionary.org.

[34] https://metapro.ruimao.tech/ Accessed 1st October 2022.

concepts were abstracted from the original metaphor and its literal counterpart, respectively. Finally, the target concept was connected to the source concept by “is” in the output. In this way, Mao et al. (2023) could achieve concept mappings for verb, noun, adjective, and adverb metaphors. However, the limitation is that the system cannot abstract concepts for MWEs.

To sum up, abstraction-driven methods could generate abstract concepts representing a group of words. Multiple metaphorical words with similar source and target domains could be connected with concept mappings. The outputs of abstraction-driven methods were the closest to CMT-argued concept mapping representations. Thus, the generated concept mappings could be directly used for understanding human cognition for metaphors and other psycho-linguistic research. However, current abstraction-driven methods could not process the concept mapping tasks for multi-word expressions.

5.4 Summary

Table 15 shows that a few conceptual metaphor processing tasks (Mason 2004; Gagliano et al. 2016) simply processed the task without metaphor identification. Most studies took word pairs as input to simplify the learning task, whereas Rosen (2018) and Mao et al. (2023) processed the task from whole sentences, covering all open-class words. Gandy et al. (2013); Strzalkowski et al. (2013); Dodge et al. (2015) used domain-specific corpora to study the task. Many related words did not aim for processing concept mappings for MWEs, besides the work of Rosen (2018).

Compared with features used in linguistic metaphor processing, Table 16 shows that lexical resources were critical for conceptual metaphor processing. PoS was also an essential feature because different PoS have different clustering, conceptualization, and abstraction paradigms in conceptual metaphor processing. Many researchers (Strzalkowski et al. 2013; Fu et al. 2020; Ge et al. 2022; Mao et al. 2023) used word co-occurrence features to learn the task because the conceptual mapping inferences largely depend on word co-occurrences of literal instances. Although conceptual metaphor processing is a cognitive learning task, few current methods employed cognitive intuitions in their features. For example, abstractness and imageability used in the works of Gandy et al. (2013); Strzalkowski et al. (2013) are cognition-related features. The former showed the property of a concept to be considered far away from a particular object or instance. The latter showed the property of a concept to evoke an image in the mind when recognizing it, respectively.

Table 17 shows common learning paradigms in conceptual metaphor processing. Many studies before 2016 employed rule-based algorithms (Gandy et al. 2013; Strzalkowski et al. 2013) and selectional preference violation (Mason 2004; Li et al. 2013). The most up-to-date research (Ge et al. 2022; Mao et al. 2023) proposed an end-to-end model based on distant supervised learning and neural networks.

Table 18 illustrates the evaluation methods and metrics in conceptual metaphor processing studies. Since concept mapping is lacking in datasets with golden labels, most studies used human evaluation as a supplement. Ge et al. (2022) additionally evaluated the quality of concept mappings by the performance improvements on the metaphor identification task. Mao et al. (2023) qualitatively evaluated their system by the levels of analysis, e.g., linguistic metaphor, conceptual metaphor, extended metaphor, metaphorical inference processing, and applicability, e.g., task coverage, easy-to-integrate, unrestricted text, and open-domain processing. The evaluation framework was proposed by Shutova (2015).

Table 15 Conceptual metaphor processing models by author, task definition, and research scope

Category	Paper	Id.	Rel.-level	Sent.-level	PSG-level	Verb	Adj.	Noun	Adv.	SD
Clustering	Mason (2004)		✓			✓				
	Gandy et al. (2013)	✓	✓			✓	✓	✓		✓
	Strzalkowski et al. (2013)	✓			✓	✓				✓
Conceptual- ization	Li et al. (2013)	✓	✓			✓				
	Gagliano et al. (2016)		✓					✓		
	Rosen (2018)			✓		✓	✓	✓	✓	
Abstraction	Dodge et al. (2015)	✓	✓			✓	✓	✓		✓
	Fu et al. (2020)							✓		✓
	Ge et al. (2022)	✓	✓			✓	✓			
	Mao et al. (2023)	✓		✓		✓	✓	✓	✓	

PSG denotes passage

Table 16 Conceptual metaphor processing models by author and feature

Category	Paper	POS	Lem.	MI	Cooc.	Sem. attr.	Lex. Res.	Word emb.	PLM	Graph emb.
Clustering	Mason (2004)	✓					✓			
	Gandy et al. (2013)	✓		✓		✓	✓			
	Strzalkowski et al. (2013)	✓		✓	✓	✓	✓			
Conceptual- ization	Li et al. (2013)	✓	✓		✓		✓			
	Gagliano et al. (2016)	✓					✓	✓		
	Rosen (2018)	✓								
Abstraction	Dodge et al. (2015)	✓	✓				✓			
	Fu et al. (2020)	✓			✓		✓	✓		✓
	Ge et al. (2022)	✓	✓		✓		✓		✓	
	Mao et al. (2023)	✓	✓		✓		✓		✓	

Lem. lemma, *MI* mutual information

Table 17 Conceptual metaphor processing models by author and learning paradigm

Category	Paper	Rule	SPV	Cos.	Unsup.	Cluster	NN	E2E	Graph-based
Clustering	Mason (2004)		✓		✓	✓			
	Gandy et al. (2013)	✓							
	Strzalkowski et al. (2013)	✓				✓			
Conceptual- ization	Li et al. (2013)	✓	✓		✓				
	Gagliano et al. (2016)			✓	✓				
	Rosen (2018)						✓		
Abstraction	Dodge et al. (2015)	✓							
	Fu et al. (2020)			✓					✓
	Ge et al. (2022)						✓	✓	
	Mao et al. (2023)						✓	✓	

E2E End-to-End

Table 18 Conceptual metaphor processing models by author and evaluation

Category	Paper	Auto.	Human	Metrics
Clustering	Mason (2004)		✓	Self-defined polarity
	Gandy et al. (2013)		✓	Validness, meaningfulness
	Strzalkowski et al. (2013)	✓	✓	Accuracy, affect, commonness
Conceptual- ization	Li et al. (2013)	✓		Match Top 1/3, correct Top 3
	Gagliano et al. (2016)			–
	Rosen (2018)	✓		Accuracy
Abstraction	Dodge et al. (2015)			-
	Fu et al. (2020)	✓		Hits@N
	Ge et al. (2022)	✓	✓	F1, source/target domain suitability
	Mao et al. (2023)		✓	Qualitative evaluation framework proposed by Shutova (2015)

6 Metaphor generation

Metaphor generation can be categorized into the following sub-tasks.

(1) Verb substitution is a task that replaces a literal verb with a metaphorical one, which is the opposite of the word-level paraphrasing-based metaphor interpretation task.
(2) Metaphor surface realization (MSR) generates properties to realize a metaphor with an implicit cognitive connection between the given source and target words. Alternatively, given a target word and a property, MSR means to generate a source word, forming a simile.
(3) Sentence generation is a task that generates metaphors given a target word.

The difference between MSR-orientated and sentence generation-orientated tasks is that the former follows a specific template or extracts a phrase from a corpus to construct a metaphor, while the latter generates a sentence via language models.

6.1 Verb substitution

Yu and Wan (2019) presented an end-to-end metaphor generation framework. They first obtained source candidate sets from synonyms and hypernyms of targets in WordNet with similarity conditions. Subsequently, they trained a PoS-constrained language model to generate a sentence with a target verb. Finally, they applied an adjustable joint beam search algorithm in the decoding phase to guarantee the metaphoricity of substituting the source verb. Their automatic evaluation used perplexity scores, while human evaluation measured readability, creativity, and metaphorical or literal verb usage for the generated sentences.

Stowe et al. (2020) proposed a lexical replacement method and a metaphor masking Seq2Seq model to avoid creating a large parallel corpus. The former method was inspired by the work of Mao et al. (2018). The candidate words were selected from hyponyms of target words. The one with the highest cosine similarity between the word embeddings of the context and itself was defined as the replacement. The latter method masked the target words in metaphorical sentences. It outputted the metaphorical sentences during training, masked the target words in literal sentences, and generated metaphorical sentences during testing. They used crowd-sourcing to evaluate the metaphoricity, fluency, and paraphrase quality of generated metaphorical sentences.

Chakrabarty et al. (2021) supposed that metaphors widely existed in poetry and could enhance poetry creativity. They fine-tuned a pre-trained Seq2Seq language generational model, BART (Lewis et al. 2020), by encoding literal sentences and decoding metaphorical sentences. They also added a metaphor discriminator to avoid introducing noise. The model performance was measured by semantic similarity, BLEU-2 (Papineni et al. 2002), and BERTScore (Zhang et al. 2019) in automatic evaluation, fluency, meaning, creativity, and metaphoricity in human evaluation.

Stowe et al. (2021a) analyzed the difference between free and controlled generation with concept mappings using a Seq2Seq T5 model (Raffel et al. 2020). The results were evaluated by automatic and human evaluation. The evaluation showed that free generations could be more fluent, while controlled models could generate more novel metaphors. Based on the evaluation results, they also tested the correlations between automatic and human evaluation metrics about metaphors. SentBERT (Reimers and Gurevych 2019) and MoverScore (Zhao et al. 2019) could capture the semantic similarity. Perplexity was the best to describe fluency. Binary metaphor classification (Su et al. 2020a) could measure metaphoricity.

Stowe et al. (2021b) proposed two metaphor generation models (CM-Lex and CM-BART) based on concept mappings from FrameNet (Baker et al. 1998). CM-Lex was an unsupervised method with a linear transformation from source and target domains to source and target words. CM-BART fine-tuned BART with concept mappings. The training dataset was similar to that of Chakrabarty et al. (2021). The testing dataset included 150 literal/metaphorical sentence pairs, selected or generated from the Gutenberg Poetry corpus, the Mohammad 2016 corpus (Mohammad et al. 2016), and the Brown Corpus (Francis and Kucera 1979) for evaluating the diversity of metaphors. The automatic evaluation used cosine distance to show the model performance. The human evaluation focused on the metaphoricity and whether the generated source word belonged to the source domain.

Verb substitution is the most widely studied sub-task in the metaphor generation domain. Firstly, verb metaphors are more frequent than metaphors in other PoS. Furthermore, verb substitution is a relatively simple task in metaphor generation because a literal verb can be replaced with a metaphorical verb to generate a metaphor without modifying the syntactic structure of the literal sentence. However, the limitation was that such a verb substitution method failed to generate a metaphor with diverse syntactic structures and language styles. It limited the creativity and novelty of a generated sentence.

6.2 Metaphor surface realization

Zheng et al. (2019) believed that metaphors could improve users' participation in human-computer conversations and proposed an unsupervised method for metaphor generation. They collected target words from poetry themes and source words from their chatbot log, filtered by concreteness and frequency. This MSR task was defined as finding a connecting word as a property shared by the source and target pair since the context had been given in their collected corpus. In this model, the connecting word was obtained by a connecting score designed with a distance function. However, the source and target pairs were selected randomly.

Song et al. (2020) transformed MSR as a knowledge graph completion task. The task would generate a source concept based on a target and an attribute. The graph embedding combined a metaphor knowledge graph with concept-attribute collocations. The generated source concept word was fitted in a simile template to construct a figurative expression.

Current MSR methods focused on nominal metaphors. Most of the MSR studies aimed to generate simile expressions. The generated sentences followed similar syntactic structures using known phrases in a corpus. However, the pre-defined phrases could not be used in open-domain metaphor generation tasks, which resulted in limited usability for current MSR systems in real-world tasks.

6.3 Sentence generation

Brooks and Youssef (2020) observed that many syntactic patterns in metaphors never appeared in literal texts. They proposed a framework starting with an unsupervised LSTM language model to generate a sentence containing words with a weighted score under constraints. Next, the model identified syntactic patterns for metaphor or literal expressions. The model would check the metaphoricity and novelty of the generated sentence with syntactic patterns during inference. The evaluation used unique syntactic patterns or sub-patterns to measure the model performance.

Table 19 Metaphor generation models by author and task definition

Category	Paper	Verb	Adjective	Noun	Multi-lingual
Verb substitution	Yu and Wan (2019)	✓			
	Stowe et al. (2020)	✓			
	Stowe et al. (2021a)	✓			
	Stowe et al. (2021b)	✓			
	Chakrabarty et al. (2021)	✓			
MSR	Zheng et al. (2019)			✓	✓
	Song et al. (2020)			✓	
Sentence generation	Brooks and Youssef (2020)	✓	✓	✓	
	Li et al. (2022)			✓	

Li et al. (2022) focused on Chinese nominal metaphor generation with a GPT2 (Radford et al. 2019)-based model. They utilized large-scale unlabelled data and applied self-training to solve the issue with data sparsity. They added an auxiliary metaphor identification task to render metaphorical parts larger weights during training, which made the model relatively focus on generating metaphorical parts. Using GPT2 made the output sentences more diverse in meanings and formats. However, during inference, the generated concept was unplanned. This made the generated metaphors aimless.

In summary, sentence generation-based methods could generate more syntax-flexible sentences than other sub-tasks in the metaphor generation domain. These methods could generate more novel and creative metaphors. However, the challenge is to generate purposeful metaphors with controllable source and target concept mappings.

6.4 Summary

Table 19 shows that the most popular sub-task in metaphor generation is to replace a metaphorical verb with a literal one (Yu and Wan 2019; Stowe et al. 2020; Chakrabarty et al. 2021). Zheng et al. (2019); Song et al. (2020) worked on relation-level MSR, outputting a property or a source concept in nominal or explanatory forms. Brooks and Youssef (2020) generated a new sentence with different syntactic patterns regardless of the PoS of a metaphorical word. Li et al. (2022) focused on generating sentences with nominal metaphors since nominal metaphors contain comparatively more straightforward sentence structures.

Table 20 shows that the concept mapping feature supported verb substitution generation because the generation task needed conceptual mapping information to guide and generate appropriate metaphors. Most studies selected Lexical resources in all three sub-tasks, such as WordNet, COMET, and MetaNet.

Table 21 shows the learning paradigms used in the metaphor generation studies. Seq2Seq is a natural choice for the generation task, which generally needs parallel datasets. Seq2Seq models in current studies were used under constraints, only replacing the verbs in the generated sentences. Stowe et al. (2020) proposed a masking framework to mitigate learning dependency from literal parallel corpora. Chakrabarty et al. (2021) utilized a masking model to generate a literal dataset as the input for the generalization model based on the metaphor dataset. Unsupervised methods played a significant part in metaphor generation due to the lack of large parallel datasets in metaphor generation tasks. Zheng et al. (2019); Stowe et al.

Table 20 Metaphor generation models by author and feature

Category	Paper	POS	Sem. attr.	Lex. res.	Word emb.	PLM	Graph emb.	Concept domain
Verb substitution	Yu and Wan (2019)	✓		✓		✓		
	Stowe et al. (2020) -lexical replacement	✓		✓	✓			
	Stowe et al. (2020) -metaphor masking	✓						
	Stowe et al. (2021a)	✓		✓		✓		✓
	Stowe et al. (2021b) -CM-LEX	✓		✓	✓			✓
	Stowe et al. (2021b) -CM-BART	✓						✓
	Chakrabarty et al. (2021)	✓		✓				
MSR	Zheng et al. (2019)	✓	✓		✓			
	Song et al. (2020)	✓		✓			✓	
Sentence generation	Brooks and Youssef (2020)	✓						
	Li et al. (2022)	✓				✓		

(2020, 2021b) utilized cosine similarity to seek metaphorical words or connecting words, given literal ones. Li et al. (2022) introduced an auxiliary metaphor identification task to regulate the metaphoricity of a generated metaphor.

Table 22 shows that most generation studies utilized both automatic and human evaluation to evaluate the performance of generation models. Both automatic and human evaluation measured creativity, metaphoricity, and fluency. BLEU and sentence BERT were typically used for measuring the creativity of paraphrasing. Stowe et al. (2021a); Li et al. (2022) used a metaphor classifier to evaluate metaphoricity. Perplexity was frequently used for measuring fluency in natural language generation tasks. Besides, the automatic evaluation methods also aimed to measure semantic similarity between two sentences by semantic distance.

Contemporary research is predominantly concerned with verifying the metaphoricity of generated expressions. Most of these studies concentrated on generating expressions that deviate from the literal meaning, regardless of the source concept. Some studies (Zheng et al. 2019; Song et al. 2020; Stowe et al. 2021b) have succeeded in generating metaphors using arbitrarily selected concept mappings or attributes. Nonetheless, the meaningfulness and appropriateness of these concept mappings or attributes for practical usage still need to be determined. In practice, one may expect a metaphor generation technique with better control of source concepts and aimed metaphorical meanings.

7 Applications

Metaphor processing techniques have close connections with many downstream tasks. They can be used for text pre-processing, feature engineering, and linguistic analysis.

(1) Text pre-processing: Mao et al. (2018); Zheng et al. (2019); Chakrabarty et al. (2021); Mao et al. (2022a) used metaphor-processing techniques, such as metaphor interpretation or generation, as a text pre-processing tool to improve downstream tasks, such as machine translation, chatbot system, poetry generation, and sentiment analysis. Generally, using the texts after metaphor pre-processing could yield better performance than using the original texts.

(2) Feature engineering: Cabot et al. (2020); Zhang et al. (2021a); Han et al. (2022) used metaphor processing techniques to generate additional features. These features delivered helpful information for downstream tasks, namely political attribute and mental illness classification tasks. Because these tasks are related to cognition, metaphor features can somewhat boost their model performance.

(3) Linguistic analysis: Prabhakaran et al. (2021); Hu and Wang (2021) studied linguistics in the political domain based on metaphors because the use of metaphors in a text can reveal fine-grained sentiment and underlying cultural tendencies (Table 23).

7.1 Text pre-processing

The community has realized that metaphors are difficult for downstream NLP task learning. The semantics of metaphors are different from that of literal ones, so they are likely to cause errors for machines in natural language understanding. For example, the literal meaning of “she devoured his novels” is not positive in sentiment analysis, whereas its metaphorical meaning is positive. Without metaphor interpretation, a sentiment classifier was likely to lead to an incorrect result for the sentence (Mao et al. 2022a). Understanding metaphors is also challenging for language learners because of cultural differences. The literal translation

Table 21 Metaphor generation models by author and learning paradigm

Category	Paper	Cos. Sim.	Unsup.	DL	Pre -trained	S2S	Graph -based	Multi -task
Verb substitution	Yu and Wan (2019)			✓		✓		
	Stowe et al. (2020) -lexical replacement	✓	✓					
	Stowe et al. (2020) -metaphor masking			✓		✓		
	Stowe et al. (2021a)			✓	✓	✓		
	Stowe et al. (2021b) -CM-LEX	✓	✓					
	Stowe et al. (2021b) -CM-BART			✓	✓	✓		
	Chakrabarty et al. (2021)			✓	✓	✓		
MSR	Zheng et al. (2019)	✓	✓					
	Song et al. (2020)			✓			✓	
Sentence generation	Brooks and Youssef (2020)		✓	✓				
	Li et al. (2022)			✓	✓			✓

Table 22 Metaphor generation models by author and evaluation

Cate.	Paper	Auto.	HMN	Cre.	Meta.	Flue.	Sem. dis.	SBT	BLEU	Ppl.	Clf.
Verb sub.	Yu and Wan (2019)	✓	✓	✓	✓					✓	
	Stowe et al. (2020)		✓		✓	✓					
	Stowe et al. (2021a)		✓		✓	✓	✓	✓	✓	✓	✓
	Stowe et al. (2021b)	✓	✓		✓		✓				
	Chakrabarty et al. (2021)	✓	✓	✓	✓	✓		✓	✓		
MSR	Zheng et al. (2019)	✓	✓	✓							
	Song et al. (2020)	✓									
Sent. gen.	Brooks and Youssef (2020)	✓									
	Li et al. (2022)	✓	✓	✓		✓				✓	✓

Underlined metrics are for human evaluation *Cre.* creativity, *Meta.* metaphoricity, *Flue.* fluency, *Sem. dis.* semantic distance, *SBT* sentence BERT, *Ppl* perplexity, *Clf* classifier

Table 23 Applications by metaphor processing technique, author, and downstream task

Category	Paper	Downstream task
Text pre-processing	Mao et al. (2018)	Machine translation
	Zheng et al. (2019)	Chatbot system
	Chakrabarty et al. (2021)	Poetry generation
	Mao et al. (2022a)	Sentiment analysis
Feature engineering	Cabot et al. (2020)	Political attribute classification
	Zhang et al. (2021a)	Mental illness detection
	Han et al. (2022)	Mental illness detection
Linguistic analysis	Prabhakaran et al. (2021)	Metaphor analysis in politics
	Hu and Wang (2021)	Metaphor analysis in politics

of the same example “she devoured his novels” does not make sense in Chinese (Mao et al. 2018).

Mao et al. (2018) evaluated the performance of their proposed metaphor identification and interpretation model on a machine translation task. The results showed that paraphrasing metaphors into their literal counterparts boosted accuracy for Google translator by 26% and Bing translator by 24% in an English-Chinese metaphor translation task. Mao (2020) further evaluated the performance of the metaphor interpretation technique in seven machine translation tasks, where the source language was also English. The target languages were German, Russian, Greek, Italian, Chinese, Thai, and Japanese. All these translation tasks showed improved performance after metaphor paraphrasing, where the gains in the target languages of Asia were higher than those in the target languages of Europe.

Zheng et al. (2019) tested the metaphor generation system in an existing social chatbot. The chatbot used metaphors in conversations in two formats, one-round and two-round metaphors. In one-round metaphors, the chatbot directly responded to the whole metaphorical sentences. In two-round metaphors, it first replied with a metaphorical sentence in a simile format, followed by an explanation in the second round. Users’ follow-up rates were 22%, 27%, and 41% for literal sentences, one-round metaphors, and two-round metaphors, respectively. Metaphors improved user engagement in social chatbot conversations on a large scale.

Chakrabarty et al. (2021) used their metaphor generation model MERMAID to edit poems. They believed that using appropriate metaphors could enhance the creativity of poems. The results showed that participants from the crowd-sourcing platform preferred 68% of poems embellished by MERMAID.

Mao et al. (2022a) evaluated the performance of MetaPro on a news headline sentiment analysis dataset. The experiments showed that MetaPro could improve the performance of state-of-the-art sentiment analysis APIs. They also showed that if the train and test sets were paraphrased first before training and evaluating a classifier, then the sentiment classifier also achieved improvement in the non-metaphorical (paraphrased) dataset.

7.2 Feature engineering

Cabot et al. (2020) proposed a multi-task learning model for political attribute classification. The auxiliary tasks were token-level metaphor detection and sentence-level emotion prediction. The main tasks were targeted at predicting the political perspective of news articles, the party affiliation of politicians, and the frame of policy issues at the document level. The results showed that using the multi-task framework with metaphor prediction enhanced the performance on all three main tasks, while emotion classification improved the performance on the political perspective and party affiliation classification tasks.

Zhang et al. (2021a) observed that many patients with mental disorders were willing to share their feelings online with metaphors. They presented a novel metaphor-informed model to detect mental disorders with accurate performance on long sentences. They extracted metaphor features, such as the number of metaphors, the portion of metaphors in sentences, and the PoS of metaphors from the sentences, and fused them in sentence representations. The results showed that this model achieved at least 3% higher F1 than baselines on detecting depression and anorexia.

Han et al. (2022) argued that using metaphorical concept mapping features could improve the performance and the explainability of a depression detection task. They proposed an explainable hierarchical attention network to retrieve depression-featured tweets and concept mappings. Thus, the retrieved concept mappings could reveal the inner world of a depressed person, because these concept mapping patterns were implicitly expressed by the subject in daily communication on social media. The authors used MetaPro (Mao et al. 2023) to obtain the concept mapping features. Their experiments showed that introducing the concept mapping features could result in higher accuracy for the depression detection task. In this work, they also demonstrated the effectiveness and parameter efficiency of their proposed hierarchical attention network encoder by comparing classical encoders, such as LSTM, BiLSTM, GRU, BiGRU, and Transformer.

7.3 Linguistic analysis

Politicians tend to behave actively to engage their constituents. Using metaphors in discourse can frame human cognitive perspectives. Some researchers explored the importance of metaphors in political discourse. Prabhakaran et al. (2021) utilized a metaphor classification model from Rei et al. (2017) and analyzed metaphors used by different gender and parties during significant political events. They concluded that the metaphors used were related to ideological leanings, that current political states and posts with metaphors could engage the audience, and that metaphorical language elicited more engagement than its literal language.

Hu and Wang (2021) compared source domains in two government reports from China and the United States, respectively, and analyzed the differences and similarities of conceptual metaphor usage. The similarity between the two countries reflected the cognition of commonsense knowledge in the political domains. For example, JOURNEY source domain metaphors were used to express that achieving the goals was a long-term process. Diverse rooting cultures mainly caused the differences. For example, the US political report uses THEATRE source domain metaphors related to US entertainment.

7.4 Summary

Table 24 shows metaphor processing techniques and functions applied in downstream tasks. The table shows that token-level metaphor identification and word-level paraphrasing were more frequently used than others in downstream applications (Mao et al. 2018, 2022a; Han et al. 2022). These techniques could improve the metaphor understanding ability from semantic aspects. On the other hand, (Zheng et al. 2019; Chakrabarty et al. 2021) used metaphor generation methods, namely verb substitution and MSR, to improve language art

Table 24 Applications by author, metaphor processing technique, and function employed in downstream tasks

Category	Paper	Id.		INTPN		CMP		Gen.	
		RL	TL	WP	EP	Clust.	Abs.	VS	MSR
Text pre-processing	Mao et al. (2018)		✓	✓					
	Zheng et al. (2019)								✓
	Chakrabarty et al. (2021)							✓	
	Mao et al. (2022a)		✓	✓	✓				
Feature engineering	Cabot et al. (2020)		✓						
	Zhang et al. (2021a)		✓						
	Han et al. (2022)		✓	✓			✓		
Linguistic analysis	Prabhakaran et al. (2021)	✓							
	Hu and Wang (2021)					✓			

RL relation-level, *TL* token-level, *WP* word-level paraphrasing, *EP* explanation pairing, *Clust.* clustering-driven methods, *Abs.* abstraction-driven methods, *VS* verb substitution and human-computer interaction. Lastly, conceptual metaphor processing techniques, such as clustering and abstractness-driven methods, have shown their utility in depression detection (Han et al. 2022) and political expression analysis (Hu and Wang 2021). These methods helped researchers achieve automatic and large-scale analysis in psychological and cognitive computation tasks.

8 Future Work

Computational metaphor processing lies at the intersection of computational linguistics and cognitive computing. Many current works largely depended on the power of deep learning to process the sub-tasks, whereas the linguistic and cognitive intuition were somewhat blurred in these models. To inspire future research, we want to highlight linguistics and cognition-informed studies in Table 25.

As seen in Table 25, concreteness, abstractness, imageability, and affect were commonly used cognitive features. Unlike embedding features and pre-trained language models, these features cannot be directly obtained from current language modeling methods and large-scale pre-training corpora. However, these features demonstrate specific relationships with computational metaphor processing. For example, concreteness and abstractness reflect the hypothesis that people are likely to use metaphors to explain abstract concepts (Tsvetkov et al. 2013; Gandy et al. 2013; Do Dinh and Gurevych 2016). Lakoff and Johnson (1980) argued that “love is not love without metaphors of magic, attraction, madness, union, nurturance, and so on.” The later concrete concepts and metaphors undoubtedly contribute to understanding the abstract concept of LOVE. Imageability is another cognitive feature. The hypothesis is that a metaphorical concept’s imageability differs from its context (Broadwell et al. 2013). Similar to SPV, it also explains the contrast between a metaphor and its context, although imageability does not represent co-occurrence, semantic or syntactic features. Affect score reflects the sentiment or emotional impact of a given word. The hypothesis is that metaphorical concepts deliver more decisive affective information than literal ones (Strzalkowski et al. 2013; Rai et al. 2019). Thus, modeling affect scores can differentiate metaphors from literal expressions from the emotional perspective. The above features and their associated hypotheses represent different motivations for studying metaphors. Studying metaphors in different contexts, such as linguistics, emotion, cognition, and psychology, would be valuable.

Cognitive and linguistic theories, namely CMT, SPV, and MIP, have guided multiple models. These theories explain the nature of metaphors. Thus, integrating theoretical findings in a computational metaphor processing model is likely to yield better performance than a general model (Mao et al. 2019). They also help study metaphors from linguistic and cognitive perspectives, such as understanding their real meanings and concept mappings. We encourage future works to integrate more linguistic and cognitive intuition in model designs rather than falling into a similar learning paradigm as other natural language processing tasks, such as sequence labeling tasks.

Though previous works have made great efforts in data annotation in recent years, some issues still need to be solved. How to maintain a consistent abstraction level when annotating concept mappings for metaphors and how to annotate cross-lingual data have yet to be discussed extensively. Since abstractness may be studied more in linguistics (Borghi and Zarcone 2016), seeking help from other related linguistic theories can be a possible solution for concept mapping annotation. Annotating cross-lingual data needs to pay attention to the annotators with similar backgrounds in multiple languages.

Table 25 Linguistics and cognition informed studies by author

Task	Paper	Concr./Abst.	Imagea.	Affect	CMT	SPV	MIP
DS	Mohler et al. (2016)			✓	✓		
Id.	Tsvetkov et al. (2013)	✓					
	Mohler et al. (2013)				✓		
	Tsvetkov et al. (2014)	✓	✓				
	Do Dinh and Gurevych (2016)	✓					
	Mao et al. (2019)					✓	✓
	Gong et al. (2020)	✓					
	Choi et al. (2021)					✓	✓
	Lin et al. (2021)						✓
	Su et al. (2021b)					✓	✓
	Ottolina et al. (2021)					✓	✓
	Qin and Zhao (2021)						✓
	Su et al. (2021a)	✓					
	Song et al. (2021)						✓
INTPN	Martin (1990)	✓					
	Shutova (2010)					✓	
	Rai et al. (2019)			✓			
CMP	Gandy et al. (2013)	✓					
	Strzalkowski et al. (2013)		✓	✓			
	Gagliano et al. (2016)	✓					
	Ge et al. (2022)				✓		
Gen.	Zheng et al. (2019)	✓					
	Stowe et al. (2021a)	✓			✓		
	Stowe et al. (2021b)				✓		
App.	Hu and Wang (2021)				✓		
	Han et al. (2022)				✓		

DS dataset, *App.* application, *Abst.* abstractness, *Imagea.* imageability

Linguistic metaphor identification has been widely studied with the help of two shared tasks (Leong et al. 2018, 2020) and the large-scale annotated VUA dataset (Steen et al. 2010b). Researchers have also noticed the connection between linguistic metaphor processing and other tasks such as affective computing (Xing et al. 2020; Duong et al. 2022; Mao et al. 2022b; Cambria et al. 2022a; Ma et al. 2023). However, sub-types of linguistic metaphors, such as extended metaphors and metaphorical MWEs, still need to be studied in depth. Extended metaphors exist in multiple sentences, paragraphs, or discourses with a continuous and intense comparison between source and target domains. Learning and understanding the long-term dependency on source and target domains is particularly challenging. However, it reflects the intelligence of humans in understanding complex concepts and high-level pragmatics. Fusing logical rules for metaphor understanding and reasoning is a possible direction for learning extended metaphors (Lin et al. 2023). Understanding metaphorical MWEs is also challenging. There has yet to be a method to fuse the meaning of an MWE naturally in a metaphor. It will significantly improve the usability of computational metaphor processing in downstream tasks. A concept parser is potentially helpful in understanding metaphorical concepts with MWEs. It can parse an MWE as a node in a knowledge graph to represent the conceptual relations with other nodes (Cambria et al. 2022b).

Finally, metaphor generation as an emerging task has grown fast in recent years. Generating metaphors should consider both novelty and intention. A metaphor generation system with a purposeful application scenario and controllable source and target domains would be more beneficial than randomly paraphrasing literal expressions into metaphorical ones.

9 Conclusion

Computational metaphor processing is essential in the NLP community, given how frequently it has been used in daily language. Metaphors help to enrich language art and frame human cognitive systems. Recently, researchers have explored several sub-tasks in this domain: metaphor identification, interpretation, generation, conceptual metaphor processing, and downstream task applications. These sub-tasks also raise the impacts of computational metaphor processing techniques.

This survey summarizes recent advanced and representative studies on computational metaphor processing. We review the classical metaphor theories and frequently used datasets. We focus on task definition, applied features, learning paradigms, and evaluation setups of previous works in different sub-tasks.

In recent years, metaphor identification has developed token-level PLM-based deep learning models, which are the most favored by researchers in this community. Metaphor interpretation methods included different task formats for various syntactic patterns, such as property extraction for noun metaphors and word-level paraphrasing for verb metaphors. Explanation pairing can deal with metaphors with more complex syntactic patterns, requiring more vital representation skills. Conceptual metaphor processing methods focus on generating source and target concepts to understand metaphors cognitively. Abstraction-driven methods can automatically select concepts from knowledge bases to represent a bunch of metaphors. Similar to metaphor interpretation, various task formats of metaphor generation focused on metaphors with different syntactic patterns. Current metaphor generation methods pay more attention to verb and noun metaphors. Applications of computational metaphor processing models utilized metaphor-based representation or features to better capture the semantic information of text in downstream NLP tasks. We found that though metaphor identification has achieved significant advances, other sub-tasks of computational metaphor processing have yet to attract much attention due to the lack of annotated corpus and relevant theories. We hope this survey can provide a systematic and informative view of current progress and inspire future work in diverse domains of computational metaphor processing.

Acknowledgements 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).

Author Contributions MG and RM wrote the main manuscript and Erik Cambria reviewed the manuscript.

Declarations

Competing interest The authors declare no competing interests.

References

Agirre E, Stevenson M (2007) Knowledge sources for WSD. In: Word sense disambiguation. Springer, pp 217–251

Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: The 17th international conference on computational linguistics (COLING)

Barcelona A et al (2000) Metaphor and metonymy at the crossroads. De Gruyter Mouton, New York

Baroni M, Bernardini S, Ferraresi A et al (2009) The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Lang Resour Eval 43(3):209–226

Barsalou LW (2019) Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols. In: Theories of memory. Psychology Press, pp 29–101

Bergsma S, Lin D, Goebel R (2008) Discriminative learning of selectional preference from unlabeled text. In: Proceedings of the 2008 conference on empirical methods in natural language processing, pp 59–68

Bickerton D (1969) Prolegomena to a linguistic theory of metaphor. Found Lang 14:34–52

Billow RM (1975) A cognitive developmental study of metaphor comprehension. Dev Psychol 11(4):415

Birke J, Sarkar A (2006) A clustering approach for nearly unsupervised recognition of nonliteral language. In: 11th Conference of the European chapter of the association for computational linguistics, pp 329–336

Bizzoni Y, Lappin S (2018) Predicting human metaphor paraphrase judgments with deep neural networks. In: Proceedings of the workshop on figurative language processing, pp 45–55

Borghi AM, Zarcone E (2016) Grounding abstractness: abstract concepts and the activation of the mouth. Front Psychol 7:1498

Bosselut A, Rashkin H, Sap M, et al (2019) COMET: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317

Brants T, Franz A (2006) Web 1T 5-gram Version 1. Linguistic data consortium. Philadelphia LDC2006T13

Brinton DM, Brinton LJ (2010) The linguistic structure of modern English. pp 1–446

Broadwell GA, Boz U, Cases I, et al (2013) Using imageability and topic chaining to locate metaphors in linguistic corpora. In: International conference on social computing, behavioral-cultural modeling, and prediction, Springer, pp 102–110

Brooks J, Youssef A (2020) Discriminative pattern mining for natural language metaphor generation. In: 2020 IEEE International Conference on Big Data (Big Data), IEEE, pp 4276–4283

Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods 46(3):904–911

Bulat L, Clark S, Shutova E (2017) Modelling metaphor with attribute-based semantics. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, Short Papers, pp 523–528

Burbules NC, Schraw G, Trathen W (1989) Metaphor, idiom, and figuration. Metaphor Symb 4(2):93–110

Cabot PLH, Dankers V, Abadi D et al (2020) The pragmatics behind politics: modelling metaphor, framing and emotion in political discourse. Find Assoc Comput Linguist 2020:4479–4488

Cambria E, Liu Q, Decherchi S, et al (2022a) SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: Proceedings of the 13th conference on language resources and evaluation (LREC), pp 3829–3839

Cambria E, Mao R, Han S, et al (2022b) Sentic parser: a graph-based approach to concept extraction for sentiment analysis. In: 2022 international conference on data mining workshops (ICDMW). IEEE, Orlando, pp 413–420, https://sentic.net/sentic-parser.pdf

Chakrabarty T, Zhang X, Muresan S, et al (2021) MERMAID: metaphor generation with symbolism and discriminative decoding. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4250–4261

Chen X, Leong CW, Flor M, et al (2020) Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In: Proceedings of the second workshop on figurative language processing, pp 235–243

Chen X, Hai Z, Wang S et al (2021) Metaphor identification: a contextual inconsistency based neural sequence labeling approach. Neurocomputing 428:268–279

Choi M, Lee S, Choi E, et al (2021) MelBERT: metaphor detection via contextualized late interaction using metaphorical identification theories. In: 2021 conference of the North American chapter of the association for computational linguistics: human language technologies

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46

Consortium B (2007) British national corpus. Oxford Text Archive Core Collection, Oxford

Consortium LD (2011) Spanish Gigaword, 3rd edn. Linguistic Data Consortium

Do Dinh EL, Gurevych I (2016) Token-level metaphor detection using neural networks. In: Proceedings of the fourth workshop on metaphor in NLP, pp 28–33

Dodge EK, Hong J, Stickles E (2015) MetaNet: Deep semantic automatic metaphor analysis. In: Proceedings of the third workshop on metaphor in NLP, pp 40–49

Duong C, Liu Q, Mao R, et al (2022) Saving earth one tweet at a time through the lens of artificial intelligence. In: 2022 International joint conference on neural networks (IJCNN), Padua, pp 1–9, https://doi.org/10.1109/IJCNN55064.2022.9892271

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

Francis WN, Kucera H (1979) Brown corpus manual. Lett Editor 5(2):7

Fu C, Wang J, Sang J, et al (2020) Beyond literal visual modeling: Understanding image metaphor based on literal-implied concept mapping. In: International conference on multimedia modeling, Springer, pp 111–123

Gagliano A, Paul E, Booten K, et al (2016) Intersecting word vectors to take figurative language to new heights. In: Proceedings of the fifth workshop on computational linguistics for literature, pp 20–31

Gallant SI (1991) A practical approach for representing context and for performing word sense disambiguation using neural networks. Neural Comput 3(3):293–309

Gandy L, Allan N, Atallah M, et al (2013) Automatic identification of conceptual metaphors with limited knowledge. In: Twenty-Seventh AAAI Conference on Artificial Intelligence

Gao G, Choi E, Choi Y, et al (2018) Neural metaphor detection in context. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 607–613

Ge M, Mao R, Cambria E (2022) Explainable metaphor identification inspired by conceptual metaphor theory. Proc AAAI Conf Artif Intell 36(10):10,681-10,689

Gong H, Gupta K, Jain A, et al (2020) IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In: Proceedings of the second workshop on figurative language processing, pp 146–153

Graff D, Cieri C (2003) English gigaword, linguistic data consortium

Gutierrez ED, Shutova E, Marghetis T, et al (2016) Literal and metaphorical senses in compositional distributional semantic models. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 183–193

Han S, Mao R, Cambria E (2022) Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In: Proceedings of the 29th international conference on computational linguistics (COLING). International committee on computational linguistics, Gyeongju, Republic of Korea, pp 94–104

Hazarika D, Poria S, Gorantla S, et al (2018) CASCADE: Contextual sarcasm detection in online discussion forums. In: Proceedings of the 27th international conference on computational linguistics, pp 1837–1848

Hu R, Wang X (2021) A cognitive pragmatic analysis of conceptual metaphor in political discourse based on text data mining. In: 2021 4th international conference on information systems and computer aided education, pp 235–238

Indurkhya B (2013) Metaphor and cognition: an interactionist approach, vol 13. Springer, New York

Jacobs AM (2018) The Gutenberg English poetry corpus: exemplary quantitative narrative analyses. Front Digital Hum 5:5

Kehat G, Pustejovsky J (2021) Neural metaphor detection with visibility embeddings. In: Proceedings of* SEM 2021: the tenth joint conference on lexical and computational semantics, pp 222–228

Kenton JDMWC, Toutanova LK (2019) 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 (NAACL-HLT 2019), pp 4171–4186

Klebanov BB, Leong B, Heilman M, et al (2014) Different texts, same metaphors: Unigrams and beyond. In: Proceedings of the second workshop on metaphor in NLP, pp 11–17

Klebanov BB, Leong CW, Gutierrez ED, et al (2016) Semantic classifications for detection of verb metaphors. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers), pp 101–106

Klebanov BB, Leong CW, Flor M (2018) A corpus of non-native written English annotated for metaphor. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers), pp 86–91

Kovecses Z (2010) Metaphor: a practical introduction. Oxford University Press, Oxford

Krippendorff K (2011) Computing Krippendorff's alpha-reliability. University of Pennsylvania, Technical report

Krishnakumaran S, Zhu X (2007) Hunting elusive metaphors using lexical resources. In: Proceedings of the workshop on computational approaches to figurative language, pp 13–20

Lakoff G (1994) Master metaphor list. University of California, California

Lakoff G, Johnson M (1980) Metaphors we live by. University of Chicago Press, Chicago

Le D, Thai M, Nguyen T (2020) Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In: Proceedings of the AAAI conference on artificial intelligence, pp 8139–8146

Leong CW, Klebanov BB, Shutova E (2018) A report on the 2018 VUA metaphor detection shared task. In: Proceedings of the workshop on figurative language processing, pp 56–66

Leong CW, Klebanov BB, Hamill C, et al (2020) A report on the 2020 VUA and TOEFL metaphor detection shared task. In: Proceedings of the second workshop on figurative language processing, pp 18–29

Levesque H, Davis E, Morgenstern L (2012) The Winograd schema challenge. In: Thirteenth international conference on the principles of knowledge representation and reasoning

Lewis M, Liu Y, Goyal N, et al (2020) 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, pp 7871–7880

Li H, Zhu KQ, Wang H (2013) Data-driven metaphor recognition and explanation. Trans Assoc Comput Linguist 1:379–390

Li L, Sporleder C (2009) Classifier combination for contextual idiom detection without labelled data. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp 315–323

Li S, Yang L, He W, et al (2021) Label-enhanced hierarchical contextualized representation for sequential metaphor identification. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 3533–3543

Li Y, Lin C, Guerin F (2022) CM-Gen: A neural framework for Chinese metaphor generation with explicit context modelling. In: Proceedings of the 29th international conference on computational linguistics, pp 6468–6479

Lin Q, Mao R, Liu J et al (2023) Fusing topology contexts and logical rules in language models for knowledge graph completion. Inform Fusion 90:253–264

Lin Z, Ma Q, Yan J, et al (2021) CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 3888–3898

Liu E, Cui C, Zheng K, et al (2022) Testing the ability of language models to interpret figurative language. arXiv e-prints pp arXiv-2204

Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv e-prints pp arXiv-1907

Lönneker-Rodman B (2008) The Hamburg metaphor database project: issues in resource creation. Lang Resour Eval 42(3):293–318

Ma Y, Mao R, Lin Q et al (2023) Multi-source aggregated classification for stock price movement prediction. Inform Fus 91:515–528

Mao R (2020) Computational metaphor processing. PhD thesis, University of Aberdeen, Scotland

Mao R, Li X (2021) 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, pp 13534–13542

Mao R, Lin C, Guerin F (2018) Word embedding and WordNet based metaphor identification and interpretation. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 1222–1231

Mao R, Lin C, Guerin F (2019) End-to-end sequential metaphor identification inspired by linguistic theories. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 3888–3898

Mao R, Li X, Ge M et al (2022) MetaPro: a computational metaphor processing model for text pre-processing. Inform Fus 86–87:30–43

Mao R, Liu Q, He K, et al (2022b) The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. IEEE transactions on affective computing

Mao R, Li X, He K, et al (2023) MetaPro online: a computational metaphor processing online system. In: Proceedings of the 61th annual meeting of the association for computational linguistics (system demonstrations), pp 127–135

Martin JH (1990) A computational model of metaphor interpretation. Academic Press Professional, Inc, Boston

Mason ZJ (2004) CorMet: a computational, corpus-based conventional metaphor extraction system. Comput Linguist 30(1):23–44

Mikolov T, Sutskever I, Chen K et al (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inform Process Syst 89:59

Miller GA (1998) WordNet: an electronic lexical database. MIT Press, Cambridge

Mohammad S, Shutova E, Turney P (2016) Metaphor as a medium for emotion: an empirical study. In: Proceedings of the fifth joint conference on lexical and computational semantics, pp 23–33

Mohler M, Bracewell D, Tomlinson M, et al (2013) Semantic signatures for example-based linguistic metaphor detection. In: Proceedings of the first workshop on metaphor in NLP, pp 27–35

Mohler M, Brunson M, Rink B, et al (2016) Introducing the LCC metaphor datasets. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 4221–4227

Osbeck LM, Nersessian NJ (2010) Science as psychology: sense-making and identity in science practice. Cambridge University Press, Cambridge

Ottolina G, Palmonari M, Alam M, et al (2021) On the impact of temporal representations on metaphor detection. arXiv preprint arXiv:2111.03320

Papineni K, Roukos S, Ward T, et al (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318

Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. arxiv preprint. arXiv preprint arXiv:1802.05365

Prabhakaran V, Rei M, Shutova E (2021) How metaphors impact political discourse: a large-scale topic-agnostic study using neural metaphor detection. arXiv preprint arXiv:2104.03928

Pragglejaz G (2007) MIP: a method for identifying metaphorically used words in discourse. Metaphor Symb 22(1):1–39

Qin W, Zhao D (2021) Background semantic information improves verbal metaphor identification. In: CCF international conference on natural language processing and Chinese computing, Springer, pp 288–300

Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Raffel C, Shazeer N, Roberts A et al (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21:1–67

Rai S, Chakraverty S (2020) A survey on computational metaphor processing. ACM Comput Surv (CSUR) 53(2):1–37

Rai S, Chakraverty S, Tayal DK et al (2019) Understanding metaphors using emotions. New Gener Comput 37(1):5–27

Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009), pp 147–155

Rei M, Bulat L, Kiela D, et al (2017) Grasping the finer point: a supervised similarity network for metaphor detection. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 1537–1546

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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), pp 3982–3992

Ren L, Xu B, Lin H et al (2021) ABML: attention-based multi-task learning for jointly humor recognition and pun detection. Soft Comput 25(22):14,109-14,118

Ritter A, Etzioni O, et al (2010) A latent Dirichlet allocation method for selectional preferences. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp 424–434

Rohanian O, Rei M, Taslimipoor S, et al (2020) Verbal multiword expressions for identification of metaphor. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 2890–2895

Rosen Z (2018) Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based construction grammatical cues. In: Proceedings of the workshop on figurative language processing, pp 102–109

Rundell M, Fox GE (2002) Macmillan English dictionary for advanced learners. Korea TESOL J 5(1):183–187

Sam G, Catrinel H (2006) On the relation between metaphor and simile: when comparison fails. Mind Lang 21(3):360–378

Satopaa V, Albrecht J, Irwin D, et al (2011) Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st International conference on distributed computing systems workshops, IEEE, pp 166–171

Schuler KK (2005) VerbNet: a broad-coverage. Comprehensive verb lexicon. University of Pennsylvania, Philadelphia

Shakespeare W (2019) As you like it. In: One-hour Shakespeare. Routledge, p 56

Sharma C, Bhageria D, Scott W, et al (2020) SemEval-2020 task 8: memotion analysis-the visuo-lingual metaphor! arXiv preprint arXiv:2008.03781

Shutova E (2010) Automatic metaphor interpretation as a paraphrasing task. In: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, pp 1029–1037

Shutova E (2015) Design and evaluation of metaphor processing systems. Comput Linguist 41(4):579–623

Shutova E, Teufel S (2010) Metaphor corpus annotated for source-target domain mappings. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10)

Shutova E, Kiela D, Maillard J (2016) Black holes and white rabbits: Metaphor identification with visual features. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 160–170

Sivakumar V, Gordo A, Paluri M (2018) Rosetta: understanding text in images and videos with machine learning. Facebook Eng Blog Posted 11:2018

Socher R, Perelygin A, Wu J, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1631–1642

Song W, Guo J, Fu R et al (2020) A knowledge graph embedding approach for metaphor processing. IEEE/ACM Trans Audio Speech Lang Process 29:406–420

Song W, Zhou S, Fu R, et al (2021) Verb metaphor detection via contextual relation learning. 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), pp 4240–4251

Steen G, Dorst L, Herrmann B, et al (2010a) A method for linguistic metaphor identification from MIP to MIPVU preface. Method for linguistic metaphor identification: from MIP To MIPVU 14:IX-+

Steen GJ, Dorst AG, Herrmann JB et al (2010) Metaphor in usage. Cogn Linguist 21(4):765–796

Stemle E, Onysko A (2018) Using language learner data for metaphor detection. In: Proceedings of the workshop on figurative language processing, pp 133–138

Stowe K, Moeller S, Michaelis L, et al (2019) Linguistic analysis improves neural metaphor detection. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL), pp 362–371

Stowe K, Ribeiro L, Gurevych I (2020) Metaphoric paraphrase generation. arXiv preprint arXiv:2002.12854

Stowe K, Beck N, Gurevych I (2021a) Exploring metaphoric paraphrase generation. In: Proceedings of the 25th conference on computational natural language learning, pp 323–336

Stowe K, Chakrabarty T, Peng N, et al (2021b) Metaphor generation with conceptual mappings. 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), pp 6724–6736

Strzalkowski T, Broadwell GA, Taylor S, et al (2013) Robust extraction of metaphor from novel data. In: Proceedings of the first workshop on metaphor in NLP, pp 67–76

Su C, Huang S, Chen Y (2015) Context-dependent metaphor interpretation based on semantic relatedness. In: Natural language processing and Chinese computing. Springer, pp 182–193

Su C, Huang S, Chen Y (2017) Automatic detection and interpretation of nominal metaphor based on the theory of meaning. Neurocomputing 219:300–311

Su C, Fukumoto F, Huang X, et al (2020a) DeepMet: a reading comprehension paradigm for token-level metaphor detection. In: Proceedings of the second workshop on figurative language processing, pp 30–39

Su C, Peng Y, Huang S et al (2020) A metaphor comprehension method based on culture-related hierarchical semantic model. Neural Process Lett 51(3):2807–2826

Su C, Chen W, Fu Z et al (2021) Multimodal metaphor detection based on distinguishing concreteness. Neurocomputing 429:166–173

Su C, Wu K, Chen Y (2021) Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. Find Assoc Comput Linguist: ACL-IJCNLP 2021:1280–1287

Sweetser E (1990) From etymology to pragmatics: metaphorical and cultural aspects of semantic structure, vol 54. Cambridge University Press, Cambridge

Tasić M, Stamenković D (2015) The interplay of words and images in expressing multimodal metaphors in comics. Procedia 212:117–122

Tileagă C (2013) Political psychology: critical perspectives. Cambridge University Press, Cambridge

Tong X, Shutova E, Lewis M (2021) Recent advances in neural metaphor processing: a linguistic, cognitive and social perspective. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4673–4686

Tsvetkov Y, Mukomel E, Gershman A (2013) Cross-lingual metaphor detection using common semantic features. In: Proceedings of the first workshop on metaphor in NLP, pp 45–51

Tsvetkov Y, Boytsov L, Gershman A, et al (2014) Metaphor detection with cross-lingual model transfer. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), pp 248–258

Turbayne CM (1964) The myth of metaphor. Br J Philos Sci 571964:15

Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inform Process Syst 30:8

Wan H, Lin J, Du J et al (2021) Enhancing metaphor detection by gloss-based interpretations. Find Assoc Computa Linguist ACL-IJCNLP 2021:1971–1981

Wan M, Xing B, Su Q, et al (2020) Sensorimotor enhanced neural network for metaphor detection. In: Proceedings of the 34th pacific Asia conference on language, information and computation, pp 312–317

Wilks Y (1975) A preferential, pattern-seeking, semantics for natural language inference. Artif Intell 6(1):53–74

Wilks Y (1978) Making preferences more active. Artif Intell 11(3):197–223

Xing F, Malandri L, Zhang Y, et al (2020) Financial sentiment analysis: an investigation into common mistakes and silver bullets. In: Proceedings of the 28th international conference on computational linguistics (COLING), pp 978–987

Xu B, Li T, Zheng J, et al (2022) MET-Meme: a multimodal meme dataset rich in metaphors. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 2887–2899

Yang L, Zeng J, Li S, et al (2021) Metaphor recognition and analysis via data augmentation. In: CCF international conference on natural language processing and Chinese computing, Springer, pp 746–757

Yu Z, Wan X (2019) How to avoid sentences spelling boring? Towards a neural approach to unsupervised metaphor generation. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp 861–871

Zayed O, McCrae JP, Buitelaar P (2018) Phrase-level metaphor identification using distributed representations of word meaning. In: proceedings of the workshop on figurative language processing, pp 81–90

Zayed O, McCrae JP, Buitelaar P (2019) Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In: 2nd conference on language, data and knowledge (LDK 2019), Schloss Dagstuhl–Leibniz–Zentrum fuer Informatik

Zayed O, McCrae JP, Buitelaar P (2020) Contextual modulation for relation-level metaphor identification. Find Assoc Comput Linguist EMNLP 2020:388–406

Zayed O, McCrae JP, Buitelaar P (2020b) Figure me out: a gold standard dataset for metaphor interpretation. In: Proceedings of the 12th language resources and evaluation conference, pp 5810–5819

Zhang D, Shi N, Peng C, et al (2021a) MAM: a metaphor-based approach for mental illness detection. In: International conference on computational science, Springer, pp 570–583

Zhang D, Zhang M, Zhang H, et al (2021b) MultiMET: a multimodal dataset for metaphor understanding. 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), pp 3214–3225

Zhang T, Kishore V, Wu F, et al (2019) BERTScore: evaluating text generation with BERT. In: international conference on learning representations

Zhao W, Peyrard M, Liu F, et al (2019) MoverScore: text generation evaluating with contextualized embeddings and earth mover distance. 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), pp 563–578

Zheng D, Song R, Hu T, et al (2019) "Love is as complex as math": Metaphor generation system for social chatbot. In: Workshop on Chinese lexical semantics, Springer, pp 337–347

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.