Deep Learning-Based Knowledge Injection for Metaphor Detection: A Comprehensive Review

Anonymous ACL submission

Abstract

Metaphor as an advanced cognitive modality works by extracting familiar concepts in the target domain in order to understand vague and abstract concepts in the source domain. This helps humans to quickly understand and master new domains and thus adapt to changing environments. With the continuous development of metaphor research in the natural language community, many studies using knowledge-assisted models to detect textual metaphors have emerged in recent years. Compared to not using knowledge, systems that introduce various kinds of knowledge achieve greater performance gains and reach SOTA in a recent study. Based on this, the goal of this paper is to provide a comprehensive review of research advances in the application of deep learning for knowledge injection in metaphor detection tasks. We will first systematically summarize and generalize the mainstream knowledge and knowledge injection principles. Then, the datasets, evaluation metrics, and benchmark models used in metaphor detection tasks are examined. Finally, we explore the current issues facing knowledge injection methods and provide an outlook on future research directions.

1 Introduction

Metaphors are essentially cognitive mechanisms present in the human mind used to construct conceptual frameworks (Lakoff and Wehling, 2012). This phenomenon works by extracting familiar concepts in the target domain to understand vague and abstract concepts in the source domain (Lakoff and Johnson, 2008). As an important linguistic phenomenon, automatic detection of metaphors is crucial for many practical language processing tasks, including information extraction (Tsvetkov et al., 2013), sentiment analysis (Cambria et al., 2017), machine translation (Babieno et al., 2022), and seamless human-computer interaction (Rai and Chakraverty, 2021). In the philosophical account articulated in (Maloney, 1983), metaphor comprehension involves three distinct phases: comprehension of the literal interpretation, discovery of inconsistencies with the literal interpretation, and reasoning to recover the intended non-literal interpretation. This paper focuses on the metaphor detection task, the first two stages of metaphor comprehension in (Maloney, 1983). Consider a metaphor detection task example:

He's the **miracle** of the team.

The word "miracle" in the sentence is a metaphorical usage, which conveys something extraordinary about the person's work or collaboration by associating the word "miracle" with the person's outstanding performance in the team. In deep learning, metaphor detection is the process of determining whether a target word (e.g., "miracle") is a metaphorical usage, given the target word and context.

As the research on metaphor continues to deepen, more and more types of knowledge and injection methods have been actively explored. Mao et al. (2019) used general corpus information as the context of the words with detection. Le et al. (2020) attempted to apply dependency tree knowledge to metaphor detection by constructing a graph network adjacency matrix to utilize the dependency tree structural information. Su et al. (2020) used a prompting approach to transform metaphor detection to reading comprehension and introduced local textual information. Choi et al. (2021) took into account the basic and contextual meanings of target words. Recently, Zhang and Liu (2023) successfully reached the state-of-the-art of the current metaphor detection task by introducing adversarial learning and multi-task learning. The success of these studies highlights the important role of knowledge injection in systematic detection of metaphors. Compared to not using knowledge, systems that introduced various kinds of knowledge realized greater performance improvements. There-

fore, we believe it is necessary to comprehensively review and summarize the current metaphor detection systems as well as the contents and approaches of knowledge injection, in order to provide theoretical foundations and model references for future researchers engaged in the exploration of new knowledge.

Although several research surveys on metaphor detection have existed in the past. Rai and Chakraverty (2021); Ptiček and Dobša (2023) provided an overview of metaphor theory and computational processing methods, Abulaish et al. (2020) surveyed six technical approaches to metaphorical language, and Tong et al. (2021) delved into metaphor processing methods and their applications. However, none of these surveys has taken the principle of knowledge infusion as a primary research focus. Against this background, our surveys aim to fill the gap in this research area. First, we systematically sorted out the mainstream knowledge methods and knowledge injection principles, and used an innovative categorization method to organically integrate these studies. Second, we conducted an exhaustive review and analysis of the current major metaphor datasets, including their different variants, assessment metrics, and benchmarks. Finally, we provided insights into the strengths and limitations of different knowledge injection methods, and offered suggestions and outlooks for future metaphor detection research.

2 Knowledge

In this section, we provide an introduction to the types of knowledge that are commonly used in metaphor detection task and how they are used.

2.1 Syntactic Knowledge

Part-of-Speech Tagging. Part-of-Speech (POS) is the tagging of each word in a sentence to indicate its grammatical role or lexical category in the context. Commonly used POS tag sets include Universal POS tag sets (Petrov et al., 2011), which defined a simplified set of lexical tokens with 17 tokens, such as NOUN (noun), VERB (verb), and Treebank tag sets (Santorini, 1990), which had more detailed tokens, including JJ (adjective), JJS (adjective with a supreme ending -est), etc. In metaphor detection task, researchers usually combine POS knowledge directly into the input sequence (Song et al., 2021; Feng and Ma, 2022), or construct multitask learning with POS as an auxiliary task (Le et al., 2020).

Dependency Tree. A Dependency Tree (DT) is a syntactic structural tree used to efficiently represent dependency relationships between words in a sentence. In a Dependency Tree, each word is given a node and is connected by edges to represent the directional relationship from the dependent (subordinate) word to its main dependent (head of the subordinate) word. In metaphor detection task, researchers often utilize dependency tree knowledge to improve the syntactic comprehension of their models. Le et al. (2020) employed Graph Convolutional Network (GCN), which used the dependency tree knowledge as an adjacency matrix to build a graphical structure of dependency relationships between words. Some studies (Song et al., 2021; Feng and Ma, 2022), on the other hand, focused on extracting subject-verb-object relationships in dependency trees to aid in metaphor detection. Song et al. (2021) processed the output of subject-verb-object correspondences in text by combining, averaging, and maximizing to further capture the associations between structural semantics, while Feng and Ma (2022) used a BERT Decoder (Devlin et al., 2019) to allow the model to generate the start and end positions of subject-predicate-objects based on the context.

2.2 Semantic Knowledge

VerbNet. VerbNet (Schuler, 2005) is a verb categorization database containing nearly 4,000 English verb lemmas, and its category design refers to the study of Levin (Somers, 1994). In VerbNet, each verb is attributed to one or more categories that describe the semantic roles of the verb, syntactic constraints, and semantic relations between different categories, etc. VerbNet provides two main categorization approaches: based on syntactic structure and based on predicate meaning. In the metaphor detection task, researchers (Gong et al., 2020; Beigman Klebanov et al., 2016) used VerbNet's class information to convert each lexical unit into a binary feature vector.

FrameNet. The main goal of FrameNet (Baker et al., 1998; Lowe, 1997) is to provide sentences with semantic and syntactic annotations for a large part of the vocabulary in contemporary English. The corpus of this resource is built on The British National Corpus (Consortium et al., 2007).

FrameNet employs a semantic description based on frames, each of which represents a semantic concept and describes the events, participants, attributes, relations, etc. associated with that concept. The project (Fillmore et al., 2002) is an extended version of FrameNet, which adds the US National Corpus resources. In the metaphor detection, Li et al. (2023b) used the FrameNet provided by (Fillmore et al., 2002) in the task for frame prediction of target and contextual lexical units, and the prediction results will aid in metaphorical analysis.

WordNet. WordNet (Miller, 1995; Fellbaum, 1998) is a hierarchically structured lexical database in which each word forms links with other related words to represent the semantic connections between them. In the metaphor detection task, Gong et al. (2020); Beigman Klebanov et al. (2016) classified words into fifteen categories based on the semantic links between words in WordNet and converted these categories into binary feature vectors. Such feature vectors can be used to assist the metaphor detection and improve the performance of the model. And Zhang and Liu (2023) considered the first of the WordNet example sentences as literal meanings and used it for multi-task learning.

Dictionary Knowledge. Dictionary example sentences or paraphrase knowledge are intended to provide the model with knowledge of the polysemous and metaphorical meanings of the words to be detected, and help the model better understand the semantic changes and metaphorical expressions of the words to be detected in different contexts. In metaphor detection, some researchers have utilized lexical examples to extract the context-based basic meanings of the words to be detected (Zhang and Liu, 2023), instead of the traditional approach of directly using the words to be detected as the basic meanings. Su et al. (2021) combined the lexical paraphrase information into the model input to achieve knowledge fusion.

Concreteness. Concreteness is the degree to which a word is characterized by the meaning it expresses in a language. In metaphor detection, researchers often relyed on the word specificity rating dataset (Brysbaert et al., 2014). This dataset was based on the SUBTLEX-US corpus (Brysbaert and New, 2009) and covers 37,058 token-level samples. This dataset was rated using a 5-point scale from abstract to concrete, and the data was collected with the help of Internet crowdsourcing. In previous studies (Klebanov et al., 2014; Gong et al., 2020; Beigman Klebanov et al., 2016), the lexical units to be detected were transformed into binary feature vectors depending on their specificity ratings.

Topic. Using the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), research scholars extracted a model containing 100 topics from the New York Times (NYT) corpus (Sandhaus, 2008) to characterize general topics discussed by the public. In the metaphor detection task, previous research work (Klebanov et al., 2014; Gong et al., 2020; Beigman Klebanov et al., 2016) matched and associated the words in each instance with these 100 topics, followed by the calculation of probability scores for each word under each topic.

2.3 Emotional Knowledge

VAD Model. VAD (Mehrabian, 1996) is an affective classification system for describing and measuring the three main dimensions of human affective experience: valence, arousal, and dominance. EmoBank corpus (Buechel and Hahn, 2017) is a VAD model-based and balanced multi-type 10k English corpus of sentences, each labeled with one to five ratings on the three VAD dimensions. In the metaphor detection task, Dankers et al. (2019) introduced the EmoBank corpus (Buechel and Hahn, 2017) as an auxiliary task. Sentence-level sentiment regression was constructed based on each dimension in EmoBank. In its training process, a batch of metaphor or sentiment task data sampling is randomly selected for training at each step.

Hyperbole Corpus. Exaggeration usually involves over- or under-exaggeration of an emotion, sentiment or attitude. Combining a dataset for hyperbole detection with a metaphor detection task can make the model more sensitive to capturing emotions and sentiments in text. In a previous research, Badathala et al. (2023) introduced two hyperbole corpora, named HYPO and HYPO-L, and subsequently labeled them with metaphors. The results showed that multitask learning based on hyperbole and metaphor gains in both two-way performance.

3 Method

This section will comprehensively introduce the current mainstream knowledge injection methods. Table 1 demonstrates a summary of knowledge injection-based metaphor detection systems.

SK	SYK	EK	Injection Method	Core Structure	Papers
✓			Output Modulation	$V_{S,k} = f_b(S)[k], V_k = f_b(w_k), V_S = f_b(S)[0]$	(Choi et al., 2021) (Li et al., 2023b) (Wang et al., 2023) (Zhang and Liu)
			Determine whether the target word has a literal meaning shift in context. $V_{S,k}$ and V_k are the contextual and underlying meanings, respectively, and V_s is the contextual.		
	✓		Output Modulation	$h'_i = \lambda_i \sum h_i$	(Song et al., 2021) (Wang et al., 2023)
			Assign weight λ_i to the ith output token h_i based on the knowledge of dependency tree.		
	✓		Output Modulation	$h'_t = f_d(w'_{x<t}, H),\ t \in [1, 7]$	(Feng and Ma, 2022)
			Let the model generate the position h'_t of the corresponding subject and object of the target word in the sentence, and $w'_{x<t}$ for the already predicted results.		
✓	✓		Additional Inputs	$S = [cls]s_0 \oplus s_1 \oplus ... \oplus s_k[seq]$	(Gong et al., 2020) (Su et al., 2020) (Su et al., 2021)
			Combined text and knowledge input. $\oplus$ is the concat operation, $s_i(0 \leq i \leq k)$ is the context or knowledge.		
✓			Additional Inputs	$S = [cls]s_0[seq], \quad S' = [cls]s_1[seq]$	(Zhang and Liu) (Babieno et al., 2022) (Li et al., 2023a)
			Text and knowledge will be input to different Encoders. where s_0 is the input text and s_1 is the example sentence or word paraphrase.		
✓		✓	Multi-task learning	$H, Z = G(f_b(s), f_b(s'))$	(Le et al., 2020) (Mao and Li, 2021) (Mao et al., 2022) (Li et al., 2023b) (Zhang and Liu, 2023)
			Multiple task samples are used and models share parameters. H, Z correspond to the main task and subtask sample outputs, respectively, and G(x) is the fully connected layer, GCN, or gated network.		
✓			Multi-task learning	$H_0^j = f_b(g_k^j)[0]$	(Wan et al., 2021)
			A WSD subtask is designed to predict the correct meaning of the target word on the input text. g_k^j is the jth meaning of target word w_k		

Table 1: Abstract of metaphor detection system based on knowledge injection. SK: semantic knowledge. SYK: syntactic knowledge. EK: emotional knowledge. core structure: subject model architecture.

3.1 Output Modulation

Pre-defined knowledge can structure the tuning of model output to direct its attention to specific semantic content or syntactic structures. In many studies (Mao et al., 2019; Choi et al., 2021; Li et al., 2023b; Wang et al., 2023; Zhang and Liu), the main output modulation methods used are Metaphor Identification Program (MIP) and Selection Preference Violation (SPV).

MIP (Metaphor Identification Program) was originally introduced by Group (2007). Its core logic consists in comparing the difference between a lexical unit in its original meaning and its meaning in context. For the text input $S = ([cls]w_0, ..., w_k, ...w_n[seq])$ and the test word w_k, the contextual and underlying meanings of the lexical units are defined, respectively:

$$\begin{aligned} V_{S,k} &= f_b(S)[k] \\ V_k &= f_b(w_k), \end{aligned} \tag{1}$$

where f_b denotes the Encoder and $V_{S,k}$ is the corresponding output of the kth hidden layer in the text to be detected.

SPV (Selectional Preference Violation) was originally introduced to the field of metaphor detection by Wilks et al. (2013), and its core logic lies in comparing the semantic differences between the contextual meaning of a word and its surrounding words. The semantic information of surrounding words can be defined as:

$$V_S = f_b(S)[0], \tag{2}$$

where f_b denotes the Encoder and V_S denotes the corresponding hidden layer output of cls in the text to be detected. And the semantic information of SPV lexical units is similar to MIP as $V_{S,t}$.

In addition, Wang et al. (2023); Song et al. (2021); Feng and Ma (2022) apply dependency tree knowledge to adjust the outputs. In contrast, (Wang et al., 2023; Song et al., 2021) assigns different weights to the model's outputs based on the dependency tree. For the output feature $H = (h_1, h_2, \ldots, h_n)$:

$$h'_i = \lambda_i \sum h_i, \tag{3}$$

where λ_i is the corresponding weight of the ith token hidden layer output. In (Wang et al., 2023), the weight λ_i denotes the reciprocal of the distance between the ith context word and the target word. In contrast, Song et al. (2021) only focuses on the subject, object and underlying meaning of the target word, i.e., the subject and object of the target verb and the target word itself correspond to a weight of 1, while the rest of the contexts have a weight of 0.

Feng and Ma (2022) also considered the subject and object of the target word, but allowed the model to predict the position of the subject-predicate-object in the sentence:

$$h'_t = f_d(w'_{x<t}, H),\ t \in [1, 7], \tag{4}$$

where $w'_{x<t}$ is the already predicted result. h'_t is the tth predicted output, $t \in [1, 6]$ corresponds to the indexes of the beginning and the end of the subject-predicate-object in sentence H, respectively, and h'_7 is the result of metaphorical classification.

3.2 Additional Inputs

This type of approach aims to enhance the model's understanding of the context by feeding knowledge into the model along with the text to be detected. In the metaphor detection task, the researcher introduces example sentences (Zhang and Liu; Li et al., 2023a), word paraphrases (Su et al., 2021; Babieno et al., 2022), or other knowledge (Gong et al., 2020; Su et al., 2020) as additional inputs of knowledge. For input S, Su et al. (2021); Gong et al. (2020); Su et al. (2020) combines the knowledge to the input:

$$S = [cls]s_0 \oplus s_1 \oplus ... \oplus s_k[seq], \tag{5}$$

where s_0 is the input text, $\oplus$ is the concat operation, and $s_i (0 < i \leq k)$ is k paraphrases or other knowledge. And Zhang and Liu; Li et al. (2023a); Babieno et al. (2022) inputs the sentence to be detected separately from the knowledge:

$$S = [cls]s_0[seq], \quad S' = [cls]s_1[seq], \tag{6}$$

where s_0, s_1 are the input text and knowledge, respectively.

3.3 Multi-task Learning

Introducing other associated tasks can effectively promote knowledge fusion between tasks, thus helping to improve metaphor detection performance. For any metaphor and subtask sample input $s = (w_0, ..., w_n), s' = (w'_0, ..., w'_n)$, Le et al. (2020); Zhang and Liu (2023); Mao and Li (2021); Mao et al. (2022); Badathala et al. (2023); Li et al. (2023b) attempts to fuse different task knowledge, which has:

$$H, Z = G(f_b(s), f_b(s')), \tag{7}$$

where $G(x)$ denotes the layer transformation function, which can be a fully connected layer, a GCN, or a gated network. f_b is the Encoder, and H and Z denote the corresponding hidden layer outputs of the metaphorical task and other subtask samples, respectively. They are distinguished by the fact that similarity is used in (Le et al., 2020) for information alignment, with the loss defined as:

$$\mathcal{L} = \lambda ||H - Z||, \tag{8}$$

the rest label the results directly:

$$\begin{aligned} \mathcal{L}'_1 &= L_{CE}(\hat{y}_1, d_1) \quad \hat{y}_1 \in H \\ \mathcal{L}'_2 &= L_{CE}(\hat{y}_2, d_2) \quad \hat{y}_2 \in Z, \end{aligned} \tag{9}$$

where d_1 and d_2 are the true labels of the two tasks, respectively. L'_1 is a metaphor labeling task, while the definition of L'_2 subtasks varies across studies. L'_2 can be defined as a WSD task (Zhang and Liu, 2023), a POS labeling task (Mao and Li, 2021; Mao et al., 2022), an emotion labeling task (Dankers et al., 2019), an exaggeration labeling task (Badathala et al., 2023) or a FrameNet labeling task (Li et al., 2023b).

(Wan et al., 2021) fine-tunes another Encoder for any metaphorical sample $s = (w_0, ..., w_k, ..., w_n)$ and the target word w_k, g^j_k is the jth meaning of w_k, with:

$$H^j_0 = f_b(g^j_k)[0], \tag{10}$$

where H^j_0 is the jth meaning of the target word w_k corresponding to the CLS hidden layer output. If g^j_k is the contextual meaning of the target word w_k, it is labeled 1, and vice versa 0.

4 Metrics and Dataset

The purpose of this section is to provide an overview of the currently dominant metaphor detection datasets, about which detailed information has been presented in Table 3. The datasets will be presented in the following section. At the same time, we will introduce commonly used evaluation metrics in the field. In addition, we will summarize the performance of metaphor detection tasks performed on four datasets, namely, VUA ALL, VUA

	VUA ALL				VUA Verb				MOH-X (10 fold)				TroFi (10 fold)			
	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.
(Gao et al., 2018)					53.4	65.6	58.9	69.1	75.3	84.3	79.1	78.5	68.7	74.6	72	73.7
(Gao et al., 2018)	71.6	73.6	72.6	93.1	68.2	71.3	69.7	81.4	79.1	73.5	75.6	77.2	70.1	71.6	71.1	74.6
(Gao et al., 2018)	71.5	71.9	71.7	92.9	66.7	71.5	69	80.7	75.1	81.8	78.2	78.1	70.3	67.1	68.7	73.4
(Mao et al., 2019)	71.8	76.3	74	93.6	69.3	72.3	70.8	82.1	79.7	79.8	79.8	79.7	67.4	77.8	72.2	74.9
(Mao et al., 2019)	73	75.7	74.3	93.8	66.3	75.2	70.5	81.8	77.5	83.1	80	79.8	68.6	76.8	72.4	75.2
(Gong et al., 2020)	74.6	71.5	73		76.7	77.2	77						72.6	67.5	69	
(Le et al., 2020)	74.8	75.5	75.1		72.5	70.9	71.7	83.2	79.7	80.5	79.6	79.9	73.1	73.6	73.2	76.4
(Rohanian et al., 2020)									80	80.4	80.2	80.5	73.8	71.8	72.8	73.5
(Leong et al., 2020)	80.4	74.9	77.5		79.2	69.8	74.2									
(Su et al., 2020)	82	71.3	76.3		79.5	70.8	74.9		79.9†	76.5†	77.9†		53.7†	72.9†	61.7†	
(Song et al., 2021)	82.7	72.5	77.2	94.7	80.8	71.5	75.9	86.4	80	85.1	82.1	81.9	70.4	74.3	72.2	75.1
(Wan et al., 2021)	82.5	72.5	77.2	94.7	78.9	70.9	74.7	85.4								
(Choi et al., 2021)	80.1	76.9	78.5		78.7	72.9	75.7		79.3†	79.7†	79.2†		53.4†	74.1†	62†	
(Li et al., 2023b)	82.7	75.3	78.8						83.2†	84.4†	83.8†		70.7†	78.2†	74.2†	
(Babieno et al., 2022)	79.3	78.5	78.9		60.9	77.7	68.3		81	80	80.2		53.2	72.8	61.4	
(Lin et al., 2021)	79.3	78.8	79	94.8	78.1	73.2	75.6	85.8	85.7	84.6	84.7	85.2	74.4	74.8	74.5	77.7
(Wang et al., 2023)	80	78.2	79.1						77*	83.5*	80.1*		54.2*	76.2*	63.3*	
(Zhang and Liu)	80.4	78.4	79.4	94.9	78.3	73.6	75.9	86	84	84	83.4	83.6	67.5	77.6	71.9	73.6
(Feng and Ma, 2022)	81.6	77.4	79.4	95.2	81.6	71.1	76	86.4	89.5	85.2	87	87.5	72.5	77.5	74.8	77.7
(Su et al., 2021)					76	76	76	85.7	82.9	84	83.4	84.2	73.3	69.6	71.4	75.7
(Zhang and Liu, 2023)	78.4	79.5	79	94.7	78.5	78.1	78.3	87	87.4	88.8	87.9	88	70.5	79.8	74.7	76.5

Table 2: This table shows the performance of the metaphor detection system on four datasets, VUA ALL, VUA verb, MOH-X and TroFi, in recent years. Among them, most of the results on the MOH-X and TroFi datasets are based on ten-fold cross-validation, and also include some results derived from direct computation († labeling), as well as some of the models are trained on the VUA20 dataset (* labeling).

verb, MOH-X, and TroFi, in recent years, and refer to Table 2 for more details, in order to provide a comprehensive picture of the state of the art of research in this area.

4.1 Metrics

Current mainstream metaphor detection systems typically use four evaluation metrics. Among them, accuracy indicates the number of correctly categorized samples as a proportion of the total number of samples; precision measures the extent to which the model correctly predicts, focusing on the proportion of samples that the model determines to be in the positive category that are truly in the positive category; and recall measures the model's ability to correctly identify positively categorized samples (true instances). The F1-score is a metric that combines precision and recall and is used to balance the model's accuracy and recall.

4.2 Dataset

VUA. The VU Amsterdam Metaphor Corpus (Steen et al., 2010) annotates each lexical unit (187,570 in total) in a subset of the British National Corpus (Consortium et al., 2007) metaphorically. The corpus was tagged using the MIPVU metaphor detection program, and VUAMC is the largest publicly available annotated corpus of tag-level metaphor detection, and the only one to study the metaphorical nature of dummy words.

VUA SEQ. VUA SEQ is another dataset constructed based on VUAMC. Compared to VUA ALL, VUA SEQ has the same number of samples as reported (Gao et al., 2018; Neidlein et al., 2020). However, VUA SEQ covers all tokens in a sentence, even punctuation, in the classification task, thus leading to a richer number of target tokens used than VUA ALL.

VUA ALL POS. VUA ALL POS dataset has been applied to the shared task of metaphor detection (Leong et al., 2018, 2020), which consists of two parts, VUA ALL POS and VUA Verb. In particular, VUA ALL POS annotates all real-sense words (including adjectives, verbs, nouns, and adjectives) in a sentence; while VUA Verb covers only verbs. However, in the studies of (Song et al., 2021; Feng and Ma, 2022; Wan et al., 2021; Su et al., 2020), the VUA ALL POS dataset also includes dummy words. To distinguish it from the shared task (Leong et al., 2018, 2020), we named the VUA ALL POS dataset that includes both real

Dataset	#Tok.	#Sent.	%Met.
VUAall/SEQ	205,425	10,567	11.6%
VUAall/SEQ/tr	116,622	6,323	11.2%
VUAall/SEQ/val	38,628	1,550	11.6%
VUAall/SEQ/te	50,175	2,694	12.4%
VUAallpos	94,807	16,202	15.8%
VUAallpos_tr	72,611	12,122	15.2%
VUAallpos_te	22,196	4,080	17.9%
VUAverb_tr	15,516	7,479	27.9%
VUAverb_val	1,724	1,541	26.9%
VUAverb_te	5,873	2,694	29.9%
MOH-X	647	647	48.7%
TroFi	3,737	3,737	43.5%

Table 3: tr: training set. val: validation set. te: test set. tokens: number of samples. sent.: total number of sentences, %Met.: percentage of metaphorical samples

and dummy words as VUA ALL.

VUA Verb. VUA Verb is a verb part extracted from VUA (Steen et al., 2010). The number of training, validation, and test sets for VUA Verb are 15,516, 1,639, and 5,873, respectively, as reported in the metaphor detection shared task (Leong et al., 2018, 2020).

VUA18. According to (Choi et al., 2021), VUA-18 is very similar to VUA-SEQ and VUA ALL as they use the same sentences in each subset, 6,323, 1,550, and 2,694 sentences for the training, development, and test sets, respectively. VUA-18 does not consider abbreviations and punctuation as separate tokens, and has the same labeling rules as VUA ALL same as VUA ALL's labeling rules, so we group VUA-18 with VUA ALL.

VUA20. In the literature (Choi et al., 2021), VUA20 labeled 1.2k sentences with real and imaginary words. However, this does not match the description in the 20-year shared task (Leong et al., 2020). The text states that it uses the same VUA as the 18-year shared task (Leong et al., 2018) (see Section 3.1, lines 8-10) and that both report the same token count. Given this, we will not list VUA-18 and VUA-20 in the model performance table.

TroFi. TroFi (Birke and Sarkar, 2005) is a dataset focused on detecting verb metaphors, which contains the literal and metaphorical usage of 50 English verbs from the 1987-1989 Wall Street Journal corpus (Charniak et al., 2000). The dataset contains a total of 3717 samples and is not segmented.

MOH. MOH (Mohammad et al., 2016), which also focuses on verb metaphors, consists of 1639 sentences extracted from WordNet, containing 1230 sentences with literal usage and 409 metaphorical usages, which were metaphorically labeled through crowdsourcing. MOH-X (Shutova et al., 2016), on the other hand, is a subset of the MOH dataset, excluding instances with pronouns and subordinate subjects or objects, and contains 647 verb samples.

5 Future Direction

5.1 Refining the Criteria for Defining Metaphors

When introducing external knowledge, researchers often have vague criteria for defining metaphors. For example, some studies consider the first paraphrase in WordNet as the basic meaning (Zhang and Liu, 2023), or use the first k example sentences in the dictionary as the criterion for classifying non-metaphorical expressions (Su et al., 2021; Zhang and Liu). However, these approaches may introduce knowledge noise that negatively affects the performance of the model.

Precise Knowledge Annotation Methods. Future research could be conducted through manual annotation, detailed paraphrasing in specialized field dictionaries, or the use of professional vetting to ensure the accuracy and relevance of external knowledge.

Applications of Large Pre-trained Models. Utilizing large language models (e.g., GPT-3 (Brown et al., 2020)) to mine the implicit knowledge learned by the model itself. This approach can provide more accurate external knowledge for metaphor detection by analyzing the model-generated text and extracting the metaphorical information in it. For example, (Wachowiak and Gromann, 2023) guided GPT-3 to generate the target domain of metaphors.

Real-Time Knowledge Update Mechanism: Most of the past studies use older knowledge bases (Miller, 1995; Schuler, 2005; Baker et al., 1998), and it is necessary to consider designing a real-time knowledge update mechanism. By updating the external knowledge base on a regular or real-time basis to reflect changes in language usage and context, the model is better adapted to evolving contexts.

5.2 Enhancing the Knowledge Infusion Methodology

Current research has used two main approaches to inject knowledge into models: incorporating knowledge directly into the input of the model (Li

et al., 2023a; Babieno et al., 2022) or adapting the output of the model (Wang et al., 2023; Feng and Ma, 2022). However, these traditional combination approaches may not fully utilize the rich contextual information embedded in the knowledge.

Exploring More Effective Ways of Knowledge Fusion. Future research can further explore more effective ways of knowledge fusion to overcome the problem that traditional combining approaches cannot fully utilize contextual information. For example, Li et al. (2023b) introduced implicit knowledge in FrameNet by fine-tuning the model. Mao and Li (2021); Mao et al. (2022) used a gating mechanism for knowledge fusion.

Deepening Applications of Adversarial Learning. Adversarial learning Zhang and Liu (2023) has shown potential in knowledge fusion. Future research could deeply explore the use of adversarial learning to fuse other associative subtask knowledge, such as fine-grained sentiment information.

Combining Multiple Sources of Knowledge for Injection. Current research has focused on single domain knowledge injection (Badathala et al., 2023; Dankers et al., 2019), and future research could consider how to better combine multiple sources of knowledge. For example, combining hyperbole and emotion, or emotion with other rhetorical knowledge, to create more comprehensive models that are better adapted to diverse textual expressions.

5.3 Exploring Zero-shot Metaphor Detection

Most metaphorical systems are trained using manually labeled data. However, data labeling requires significant labor costs and the quality of the data greatly depends on the education level of the labeler. Zero-shot metaphor detection can alleviate the above problems to some extent.

Improvement of Existing Zero-shot Methods. In the past research, (Mao et al., 2018) introduced the cosine similarity of words, and judged words larger than a certain threshold as metaphors. And (Mao et al., 2022) defined words with the highest probability of occurrence in the BERT context as non-metaphors. The above methods have certain shortcomings in terms of detection scope and accuracy, and need to be further improved so that they can detect metaphorical expressions more accurately. For example, word-level similarity measures (e.g., cosine similarity) are extended to sentence-level or document-level similarity (e.g., WMD (Kusner et al., 2015)) to better capture the context of metaphors.

Generating Metaphor Datasets Using Large Models. Data generation with large models has been shown to be cost-effective and efficient (Wang et al., 2021; Yoo et al., 2021). Therefore, using large models to generate textual datasets containing metaphors helps to scale up the training data and reduce the reliance on manual annotation, while increasing the diversity and coverage of the data.

Guiding the Big Model for Metaphor Detection Via Prompt Learning. Prompt learning (Ye et al., 2022; Meng et al., 2022) aims to guide the LLM to generate specific content in a non-fine-tuned manner. In this task, the LLM plays the role of a few or zero sample learner. How to guide the model to deeply understand the features of metaphors and design prompts that can evoke metaphorical expressions is a very worthwhile direction for future research.

6 Conclusion

With the in-depth study of metaphor detection tasks, most models improve the detection of metaphors by injecting different knowledge. The role of knowledge injection in metaphor detection is becoming increasingly prominent. In this paper, we comprehensively review the knowledge contents and injection methods used by deep learning models on the metaphor detection task. We categorize the knowledge content into semantic, syntactic and affective knowledge. Meanwhile, we classify the knowledge injection methods as output moderation, additional input, and multi-task learning. Next, we introduce the commonly used datasets and evaluation metrics for metaphor detection tasks, and show the performance of the current metaphor detection system on four classical datasets, VUA ALL, VUAverb, TroFi and MOH-X. Finally, we discuss the problems of current metaphor detection systems and provide directions for future research.

7 Limitations

This paper provides a comprehensive description of metaphor detection systems in deep learning, focusing on discussing and summarizing in detail the different types and methods of model knowledge injection. However, there exists a small amount of research work in the area of metaphor detection that does not use knowledge or employs unsupervised methods, and these studies are not covered or

discussed in the paper. In future research, we plan to provide a comprehensive summary of most of the work in the area of metaphor detection, including both supervised and unsupervised approaches, to provide researchers with a more comprehensive understanding.

8 Ethics Statement

In this paper, we provide a detailed description of the supervised metaphor detection system and the different ways of knowledge injection. The datasets and research papers we have used have been obtained from publicly available sources and we have adhered to strict guidelines of academic and research ethics. In addition, we place special emphasis on transparency and openness of information, encourage other researchers to conduct responsible research, and uphold best practices in knowledge sharing. In the text, we explicitly cite the public data sources cited to express our full respect for the original authors and data providers of research related to the field of metaphor detection.

References

Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. 2020. A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web*, 14(1):1–52.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4):2081.

Naveen Badathala, AbisekRajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics -*.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 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)*.

Julia Birke and Anoop Sarkar. 2005. A clustering approach for nearly unsupervised recognition of nonliteral language. *Conference of the European Chapter of the Association for Computational Linguistics,Conference of the European Chapter of the Association for Computational Linguistics*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, page 977–990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. 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)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Huawen Feng and Qianli Ma. 2022. It's better to teach fishing than giving a fish: An auto-augmented structure-aware generative model for metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 656–667.

CharlesJ. Fillmore, CollinF. Baker, and Hiroaki Sato. 2002. The framenet database and software tools. *Language Resources and Evaluation,Language Resources and Evaluation*.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, page 1–39.

Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

George Lakoff and Elisabeth Wehling. 2012. *The little blue book: The essential guide to thinking and talking democratic*. Simon and Schuster.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8139–8146.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 vua metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*.

Yucheng Li, Shun Wang, Chenghua Lin, and Guerin Frank. 2023a. Metaphor detection via explicit basic meanings modelling.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loïc Barrault. 2023b. Framebert: Conceptual metaphor detection with frame embedding learning.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. Cate: A contrastive pre-trained model for metaphor detection with semi-supervised learning. *Empirical Methods in Natural Language Processing,Empirical Methods in Natural Language Processing*.

JohnB. Lowe. 1997. A frame-semantic approach to semantic annotation.

J. Christopher Maloney. 1983. A new model for metaphor. *Dialectica*, 37(4):285–301.

Rui Mao and Xiao Li. 2021. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. *Proceedings of the ... AAAI Conference on Artificial Intelligence,Proceedings of the ... AAAI Conference on Artificial Intelligence*.

Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. Metapro: A computational metaphor processing model for text pre-processing. *Information Fusion*, 86–87:30–43.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, page 261–292.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

SaifM. Mohammad, Ekaterina Shutova, and PeterD. Turney. 2016. Metaphor as a medium for emotion: An empirical study. *Joint Conference on Lexical and Computational Semantics,Joint Conference on Lexical and Computational Semantics*.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *Cornell University - arXiv,Cornell University - arXiv*.

Martina Ptiček and Jasminka Dobša. 2023. Methods of annotating and identifying metaphors in the field of natural language processing. *Future Internet*, 15(6):201.

Sunny Rai and Shampa Chakraverty. 2021. A survey on computational metaphor processing. *ACM Computing Surveys*, page 1–37.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Evan Sandhaus. 2008. The new york times annotated corpus.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision).

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 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*.

HaroldL. Somers. 1994. Book reviews: English verb classes and alternations: A preliminary investigation.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 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)*.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. A method for linguistic metaphor identification.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 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*.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing metaphor detection by gloss-based interpretations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Shun Wang, Yucheng Li, Chenghua Lin, Loïc Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Yorick Wilks, Adam Dalton, JamesF. Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Shenglong Zhang and Ying Liu. Metaphor detection via linguistics enhanced siamese network.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. *arXiv preprint arXiv:2305.16638*.