Can ChatGPT's Performance be Improved on Verb Metaphor Detection Tasks? Bootstrapping and Combining Tacit Knowledge

Anonymous ACL submission

Abstract

Metaphor detection, as an important task in the field of natural language processing, has been receiving sustained academic attention in recent years. Current research focuses on the development of supervised metaphor recognition systems, which usually require large-scale, high-quality labeled data support. With the rapid development of large-scale generative language models (e.g., ChatGPT, etc.), they have been widely used in a number of domains, including automatic summarization, sentiment analysis, and question and answer systems. However, it is worth noting that the use of ChatGPT for unsupervised metaphor detection tasks is often challenged with less-than-expected performance. Therefore, the aim of this paper is to explore how to bootstrap and combine ChatGPT by detecting the most prevalent verb metaphors among metaphors. Our approach first utilizes ChatGPT to obtain literal collocations of target verbs and subject-object pairs of verbs in the text to be detected. Subsequently, these literal collocations and subject-object pairs are mapped to the same set of topics, and finally the verb metaphors are detectd through the analysis of entailment relations. The experimental results show that the method proposed in this paper achieves the best performance on the unsupervised verb metaphor detection task compared to past unsupervised methods or direct prediction using ChatGPT.

1 Introduction

Metaphors are essentially mapping relationships between two different domains (Hesse, 1965; Lakoff and Johnson, 2008). According to Lakoff and Johnson (2008)'s theory of conceptual metaphors, linguistic metaphors derive from underlying conceptual metaphors that map a source concept to another, more abstract, domain target concept. The goal of automatic metaphor detection is to model non-literal expressions (e.g., metaphors and metonymy) and generate corresponding metaphor annotations. Improving metaphor detection is important for improving many natural language processing (NLP) tasks, including information extraction (Tsvetkov et al., 2013), sentiment analysis (Cambria et al., 2017), and machine translation (Babieno et al., 2022).

Metaphor detection, as an important part of the field of natural language processing (NLP), has seen a variety of excellent approaches emerge in recent years. In supervised classification, Su et al. (2020) constructed an ingenious framework to introduce reading comprehension into metaphor detection. Meanwhile, Choi et al. (2021) detects metaphors by comparing the basic and contextual meanings of target words. Mao and Li (2021) developed a multi-task based gating mechanism in which magnetic annotation was introduced as a secondary task. In addition, Zhang and Liu (2023) proposed a multi-task learning approach that facilitates knowledge fusion between different tasks through adversarial learning.

Supervised methods mostly rely on carefully labeled datasets, and although they show excellent performance on the corresponding test sets, they perform poorly when generalized to different domains. In the field of unsupervised metaphor detection, Heintz et al. (2013) constructed a topic table based on the latent Dirichlet allocation (LDA) (Blei et al., 2003) and aligned it to the source and target domains, respectively. While Shutova and Sun (2013) constructed a clustering map based on grammatical features of verbs, the metaphor detection system of Gandy et al. (2013) relied on lexical abstraction. Furthermore, Pramanick and Mitra (2018) calculated the abstraction levels of adjectives and nouns separately, along with the cosine distances between them, and subsequently employed the k-means algorithm for clustering. While Mao et al. (2018); Shutova et al. (2016) employed cosine similarity to determine whether the focal words belong to the same conceptual domain. Al-

though the aforementioned approaches achieved a certain level of advancement, they frequently depended on intricate manual coding rules (Heintz et al., 2013; Shutova and Sun, 2013; Gandy et al., 2013) or cannot completely escape the reliance on manually labeled datasets (Mao et al., 2018; Shutova et al., 2016).

To solve the above problems, this paper proposes an unsupervised metaphor detection method. The method explores how to bootstrap and combine ChatGPTs by detecting verb metaphors, which are the most prevalent among metaphors. First, we build a verb table that records the literal meaning collocation of each verb. Next, we introduced thematic features that map the subject and object of the target verb to one or more thematic categories. In the metaphor recognition process, we first analyze the subjects and objects of the verbs to be detected in the input text and map them to thematic categories as well. Finally, we detecte verb metaphors through the analysis of entailment relations. We conducted tests on the MOH-X and TroFi datasets, and the results show that by bootstrapping and integrating the implicit knowledge of a large language model, we can effectively improve its performance on the verb metaphor detection task.

In summary, the main contributions of this paper are summarized as follows:

1. We are the first to introduce ChatGPT to the task of verb metaphor detection and do not need to rely on tedious hand-coding rules or manually labeled data.

2. We used ChatGPT to generate a verb table that provides reference information about all literal meaning collocations for each verb.

3. We introduce topical features that act as additional semantic information to provide the method with richer background knowledge.

4. Experimental results show that by bootstrapping and integrating implicit knowledge from large language models, the method proposed in this paper achieves the best performance on the unsupervised verb metaphor detection task.

2 Related Work

The task of metaphor detection has been received a lot of attention in the field of natural language processing. Karov and Edelman (1998) used a word sense disambiguation (WSD) algorithm to cluster sentences with target words, and then made metaphor predictions based on the principle of distance between literal meanings of words. Shutova and Sun (2013) also drew on the idea of clustering, and it used the Gigaword corpus (Graff et al., 2003) with noun-related of verb-noun combinations (grammatical features) to cluster the 2000 common nouns of the BNC. In this approach, the words to be detected acquire knowledge information at a certain layer in the clustering map, i.e., the nouns at that layer are non-metaphorically related to the words to be detected.

Mao et al. (2018) presented an approximately unsupervised metaphor detection system. The system selects the best alternative to the target word by considering superlatives and synonyms in the context. When the cosine distance between the best alternative and the target word is greater than a specific threshold, it is detectd as a literal meaning. In addition, other studies Shutova et al. (2016); Pramanick and Mitra (2018) have considered the cosine distance, although Pramanick and Mitra (2018) did not use a priori labeled data to set the threshold, instead it adopted a feature construction approach using clustering for metaphorical judgments.

The studies in Turney et al. (2011); Gandy et al. (2013) explored the relationship between the abstraction degree of focus words and the expression of language metaphors. In Turney et al. (2011), the abstraction degrees of nouns, proper nouns, verbs and adverbs were first calculated, and then logistic regression was used to learn high-dimensional metaphoric features. In contrast, Gandy et al. (2013) used WordNet to generate n common collocations of the words to be detected and sorted these collocations according to the abstraction level. A metaphorical relationship word is detectd as a metaphor if it is not between the first k most concrete words. This idea is also reflected in the study of Krishnakumaran and Zhu (2007), which investigated three metaphorical relations, Subject-be-Object, Verb-Object and Adjective-Noun, and identified metaphors by determining whether the two focal words have a hyponymy relation.

Although the above methods have been effective to a certain extent, there are still problems such as complex parsing of metaphorical relationships, cumbersome construction of hand-coded knowledge, or over-reliance on manually labeled data. To overcome these challenges, this paper attempts to

introduce generative language modeling into the metaphor detection task. The main function of generative language models is to generate natural language text, which can be used for conversing with humans or performing text generation tasks. These models perform self-supervised learning from large-scale textual data without relying on task-specific labeling or guidance.

In previous research, Wachowiak and Gromann (2023) introduced generative language modeling to the field of metaphor detection for the first time, albeit with only preliminary attempts. This study first provided input text and target domain information, and then utilized ChatGPT to predict source domain information and achieved a weighted accuracy of 60.22% on the combined dataset. Inspired by this research, this paper introduces ChatGPT to the task of metaphorical sequence annotation and achieves significant performance improvements by bootstrapping and combining the model's tacit knowledge.

3 Method

In this section, we will detail the unsupervised verb metaphor recognition method by dividing its core concepts into three parts: definition of verb metaphors (§3.1), topic mapping (§3.2), and construction of verb lists (§3.3). In §3.4, we will elaborate on the specific implementation details of the proposed method.

3.1 Defining Verb Metaphors

Our study about verb metaphors is based on the theory of Selectional Preference Violation (SPV) (Wilks et al., 2013). As an important concept in linguistics, SPV reflects the relatedness and semantic compatibility between lexical units. For example, in the phrase "kill time", the verb "kill" is originally preferred to describe the behavior of animate objects, but here it modifies the inanimate "time", so there is a case of Selectional Preference Violation.

In previous studies, Shutova et al. (2012, 2016) usually categorized verb-metaphor relations into two main types, i.e., Subject-Verb (SV) pair and Verb-Direct Object (VO) pair. For example, in the sentence "He planted good ideas in their minds.", "ideas" is the direct object of the verb, and the verb "planted" forms a VO pair with "ideas". the subject of the target verb "planted" is "he", which forms an SV pair. To capture the metaphorical relations of verb pair more comprehensively, we considered both SV pair and VO pair. We consider the target verb to be non-metaphorical only if both sub-relations exhibit literal meaning relations.

In other studies, Krishnakumaran and Zhu (2007); Gandy et al. (2013) have also introduced Subject-be-Object (SbeO) relations. For example, in the sentence "Her love is a warm blanket on a cold night.", "love" is metaphorized as a warm blanket. In this structure, the verb "is" connects two focus words, "love" and "blanket". However, it should be noted that "is" as an auxiliary verb does not have an independent lexical meaning by itself; it needs to be combined with other verbs. Therefore, when judging the anaphora of SbeO structures, it is necessary to consider whether there is an entailment relationship between the subject or object. This is relatively similar to the Adjective-Noun (AN) relationship discussed in Pramanick and Mitra (2018), e.g., the SbeO structure "love is warm" with the AN structure "warm love". Therefore, we categorize SbeO relations in the same category as AN pairs instead of including them in verb metaphors.

3.2 Topic Mapping

Metaphorical relationships originated from conceptual mappings in different domains (Lakoff and Johnson, 2008). Inspired by it, we introduce the concept of topic, which can be viewed as broader and abstract concepts to correspond to domains in metaphors. Consider an example of a verb metaphor using the Oxford topic, the verb "guzzle" is often used with the subjects "baby" and the objects "milk". However, in the sentence "The car guzzled down the gasoline.", the subject and object of the target verb "guzzled" are "car" and "gasoline", respectively. This leads to the verb selective preference violation. In addition, since "bus" or "taxi" belongs to the same topic "Transport by car or lorry" as "car". Therefore, replacing the subject of the above example sentence with "bus" or "taxi" also constitutes a metaphorical expression.

We introduce three kinds of topics, namely Oxford topics, WordNet topics, and LDA topics. These three topic categories are set up in line with both the SPV (Wilks et al., 2013) and the abstractness principle defined in Turney et al. (2011); Gandy et al. (2013). The principle of abstraction holds that focus words under the same topic usually have similar or close levels of abstraction. For example, in the example in the Oxford topic, "Anger,"

Subject(Topic)	Object(Topic)
person (people)	Food or meals (Cooking and eating)
Children (Life stages)	Snacks (Cooking and eating)
Adults (Life stages)	Meat (Food)
diners (Cooking and eating)	Vegetables (Food)

Table 1: The subject and object of the verb "eat" are literally paired, with the corresponding Oxford topic category indicated in parentheses.

"Fear," and "Happiness" all belong to the "People-Feelings" topical category, and these words have similar levels of abstraction. However, it is important to note that, since a single word may have more than one denotation, the word may correspond to more than one different Oxford topic.

The LDA topics were derived from a category list containing 60 topics constructed by Heintz et al. (2013). The method first used the LDA (Blei et al., 2003) model to capture a variety of candidate topics from WiKipedia. Then, based on the metaphorical information contained in the input corpus, the topics with high relevance to metaphorical relations were selected as the final metaphorical topics, and they were summarized into 60 different topic categories. The constructed topics would be categorized according to the order of similarity in WordNet from high to low for the central words.

Similar to the infix relation defined in Krishnakumaran and Zhu (2007), we introduce the set of superlatives and synonyms in WordNet (Kilgarriff, 2000) as a third topic (WordNet topic). In WordNet, superordinates are defined as semantically more general or abstract words, while synonyms denote words with similar or identical meanings that can provide complementary information. Since both superlatives and synonyms are considered, each central word in a WordNet topic contains all synonyms and superlatives compared to LDA topics that select one or more topics by similarity.

3.3 Construction of Verb Lists

Currently, supervised metaphor detection systems (Choi et al., 2021; Zhang and Liu, 2023) usually require large-scale labeled data for training to learn the generalized distribution of metaphors. However, this data labeling process is time-consuming and labor-intensive, thus limiting its feasibility in large-scale applications. Furthermore, when supervised models are applied to transfer learning, a sharp decrease in their performance in new domains is often observed (Wang et al., 2023). This phenomenon indicates the presence of a domain bias problem, i.e., a significant difference between the metaphorical dataset and the actual metaphorical application environment. As a result, models trained on traditional datasets (e.g., TroFi or MOH-X) may have difficulty in adapting to the metaphor usage context of real application domains. To address the above challenges, we construct a verb literal meaning collocation table. This verb table requires no additional training and can be applied to detect samples with different distributions.

For the construction of verb tables, we used GPT-3.5 Turbo (hereafter Turbo) to generate literal or non-metaphorical collocations of verbs. Turbo is a lightweight text generation model developed by OpenAI that can be adapted to a variety of use cases through fine-tuning. First, we use the Turbo model to generate subject and object collocations for the target verbs (See Appendix §8.1 for details of prompt design). Then, SV and VO pairs are extracted separately by regular expressions and stored as a list. Noting that each target verb corresponds to two lists (i.e., the subject list and the object list), which do not correspond to each other. Next, we map the subject and object contents of the lists to one or more topics (see §3.2 for details), and the same topics for the same verb will be merged. Table 1 shows the Oxford topical information for the verb "eat". In the table, both "Children" and "Adult" belong to the topical category 'Life stages', so they are merged into the same category. Similarly, the object content of "Food and meals", "Snacks", "Meat" and "Vegetables" are categorized separately.

3.4 Method Implementation Details

The details of the algorithm can be found in Algorithm 1. First, we build a table of containing verbs D as described in §3.3. This verb table is in the form of a dictionary, where each particular verb is used as an indexing keyword, and the corresponding subject or object is stored in the form of a list, labeled as S_w and O_w, respectively.

To perform metaphor detection, the input text

Algorithm 1 Metaphor Detection

Require: D: Dictionary of verb forms
Require: S_w: List of literal or non-metaphorical subject topics for each target verb
Require: O_w: List of literal or non-metaphorical object topics for each target verb
Require: N: Input corpus containing sentences with target verbs
Require: w_n: Target verb in sentence n
Require: i_n: Index of the target verb in sentence n

for n in N **do**
  $S_{w_n} \leftarrow D[w_n][0]$ ▷ Retrieve subject topics
  $O_{w_n} \leftarrow D[w_n][1]$ ▷ Retrieve object topics
  Extract the **subject** and **object** from the sentence at index i_n.
  subj_nouns ← get_top_k_noun(subject)
  obj_nouns ← get_top_k_noun(object)
  subj_topics ← get_topics_from_oxford(subj_nouns)
  obj_topics ← get_topics_from_oxford(obj_nouns)
  if_sub_literal ← subj_topics $\in S_{w_n}$ ▷ Is subject literal?
  if_ob_literal ← obj_topics $\in O_{w_n}$ ▷ Is object literal?
  if $\neg$(if_sub_literal $\wedge$ if_ob_literal) **then**
    if_metaphor ← **True** ▷ Metaphor detected
  else
    if_metaphor ← **False** ▷ No metaphor
  end if
end for

needs to be processed first. Similar to the manipulation of verb lists, we will extract the subject and object in each input text. In previous studies, researchers Wilks et al. (2013); Shutova et al. (2016); Gandy et al. (2013) usually used the Stanford Dependency Parser to extract SV and VO pairs of metaphorical relations, while another study Krishnakumaran and Zhu (2007) employed PCFG (Klein and Manning, 2003) for grammatical parsing. However, these approaches usually require the specification of complex rules to take into account complex grammatical structures such as inversions, implied subjects or objects, and subordinate clauses. In this paper, we use the Turbo model to generate the subject-verb-object structure of sentences (see Appendix §8.2 for details of the prompt design). We then use regular expressions to parse the results generated by Turbo and store them as a list. If the generated SV or VO pair contain pronouns or named entities, we first obtain their basic meanings in the Oxford dictionary. For example, "it" corresponds to "used to refer to an animal or a thing that has already been mentioned or that is being talked about now". In this case, we usually choose the first 3 nouns (if they exist) as the center words of "it", such as "animal" and "thing".

Since the subjects and objects in the SV or VO pair output by the model are usually presented as phrases, we will select the first k nouns in the phrases as the center words of the subjects or objects and notate them as "subj_nouns" and "obj_nouns", respectively. Then, depending on the lexical meaning of these center words, we map them to one or more topics, denoted as "subj_topics" and "obj_topics", respectively. For example, in the sentence "He was detained on June 23, and for two weeks he was regularly assaulted by South African police", the subject of the sentence is "South African police". We extract the first k nouns as the center word, i.e., "police". According to the lexical meaning, we map "police" to the Oxford topic "Law and justice". Finally, we make metaphorical judgments based on the relationship between the parsed topics and the reference topics in the verb list.

4 Experiments

In this section, we detail the dataset used, the experimental steps, and perform an in-depth analysis of the results.

4.1 Test Datasets

To evaluate our approach, we use the MOH-X (Birke and Sarkar, 2006) and TroFi (Charniak et al.,

2000) datasets.

MOH-X. The MOH dataset was originally created by Mohammad et al. (2016), who first extracted polysemous verb samples from WordNet, and then hired 10 annotators through the crowdsourcing platform CrowdFlower3 to metaphorically annotate the sentences. To ensure the annotation quality of the dataset, Mohammad et al. (2016) used the principle of 70% annotation consistency. Furthermore, they claimed that their sample contained only two categories, literal or metaphorical, which is consistent with our hypothesis. Here, we consider only the subset of verbs (i.e., MOH-X) in the MOH dataset processed according to Shutova et al. (2016). This subset excludes instances with pronouns or subordinate subjects or objects. The dataset ultimately contained 647 verb-noun combinations, of which 316 pairs are metaphorical and 331 pairs are literal. During data preprocessing, we use a specialized tool to extract the subject-verb-object relationship of each verb to be detected and removed samples that are incorrectly parsed or lacked subjects and objects. It is worth mentioning that the MOH-X dataset we used is not further divided into a training set and a test set, but is used as a whole for model testing and evaluation.

TroFi. The TroFi dataset (Birke and Sarkar, 2006), derived from the Wall Street Journal corpus (Charniak et al., 2000), contains literal and metaphorical usage of 50 English verbs, totaling 3,717 samples, for the study of verb metaphors. Compared to the MOH-X dataset, the subject and object collocations with the target verbs in the TroFi dataset are more diverse, including pronouns, clauses, and named entities, which increases the complexity of metaphor detection. Consistent with our treatment of the MOH-X dataset, we extract subject-verb-object features for each sample in the TroFi dataset and excluded cases where parsing was wrong or where both subject and object were absent. It is worth noting that similar to the MOH-X dataset, the TroFi dataset is not further divided into training and testing sets.

4.2 Experimental Setup

Experiment 1. In Experiment 1, we compare our method with past unsupervised strong baselines (Mao et al., 2018; Shutova et al., 2016; Turney et al., 2011). Our method employs the Oxford Dictionary as the topic mapping.

In the baseline model we use, Mao et al. (2018) introduces synonyms and superlatives from WordNet, calculates the best-fit words in the input text (context) by cosine similarity, and then determines the presence of metaphors by the similarity between the fit words and the target words. In this paper, we use CBOW (Mikolov et al., 2013) for encoding and determine the similarity less than 1 as a metaphor. Unlike (Mao et al., 2018), Shutova et al. (2016) also uses cosine similarity, but only considers the similarity between the verb and the subject or object. In this paper, CBOW (Mikolov et al., 2013) is also used for encoding. If the similarity between either target word and the subject or object is less than 0, it is determined to be a metaphor. Turney et al. (2011) uses abstraction degree for metaphor judgment, again without considering the context. It assumes that relatively abstract words paired with relatively concrete words produce metaphors. In this paper, we use the abstraction degree ratings provided by (Brysbaert et al., 2014) to determine a metaphor if either target word has a relatively abstract relationship with the subject or object.

Specifically, based on the given literature descriptions, we named the above models SIM-CBOW (Mao et al., 2018), WORDCOS (Shutova et al., 2016), and Concrete-Abstract (Turney et al., 2011), respectively. in addition, we added a controlled experiment, i.e., we used GPT-3.5 Turbo directly to predict the results of the input corpus.

Experiment 2. In Experiment 2, we examined the impact of the three topic mappings introduced in §3.2 on model performance. For WordNet topics, we use the NLTK library in Python to extract the superlatives and synonyms of the central noun, and then combine all of them into the WordNet topic set corresponding to the target verb. For LDA topics, we use WUPS (Shet et al., 2012) to calculate the similarity between the central noun and the 60 LDA theme words, and classify them into one or more LDA topics based on the similarity. For Oxford topics, we first access the Oxford lexicon for pronoun disambiguation and named entity conversion, and then convert them into one or more topic categories corresponding to the Oxford lexicon.

Specifically, we first parse the input text to extract the subject and object corresponding to the target verb. We select by default the first k nouns as the subject content to be converted (k is a hyperparameter). We consider the case of extracting 1 or 3 central nouns. Specific topic types include

Models	TroFi				MOX-H			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
GPT-3.5 Turbo	58.7	11.4	64.2	19.3	60.1	20.0	91.3	32.8
SIM-CBOW	52.7	43.0	26.7	33.0	54.3	55.7	29.5	38.6
WORDCOS	54.0	44.4	38.1	41.0	59.7	74.3	26.0	38.5
Concrete-Abstract	48.9	42.2	67.2	51.8	48.3	46.2	36.9	41.1
Ours	45.8	93.7	44.2	**60.1**	61.2	93.3	56.1	**70.1**

Table 2: Performance Demonstration. both SIM-CBOW and WORDCOS are encoded using CBOW and word distances are computed using cosine similarity. Concrete-Abstract introduces lexical specificity. Our approach uses GPT-3.5 Turbo to parse verbs for literal collocations and subject-object collocations, and subsequently utilizes the Oxford Dictionary as a thematic mapping tool.

WordNet_Topic, WordNet_Topic_k, LDA_Topic, LDA_Topic_k, Oxford_Topic, Oxford_Topic_k, where k means extracting the first three nouns as the center nouns.

Experiment 3. To balance the set size with the metaphor detection accuracy when introducing topic sets, we introduce two additional hyperparameters for control. Specifically, k_1 represents the number of literal or non-metaphorical collocations selected from the verb list, while k_2 denotes the number of topics that may be covered by the subject and object corresponding to the target verb. Larger values of k_1 imply that the model's predictions cover more literal-meaning collocations of verbs, while larger values of k_2 indicate that more meanings of the subject- or object-centered words are used in the metaphorical relations parsed in the text.

To investigate the effects of the hyperparameters k_1 and k_2 on the model's metaphor detection performance, Experiment 3 was conducted accordingly. In terms of experimental design, we chose the Oxford theme. Considering the results of Experiment 2, we find that Oxford_Topic_k with three central nouns extracted performs better relative to Oxford_Topic with one central word extracted. In addition, there are relatively fewer topic types when only one central noun is extracted (which depends on the number of different meanings of that central noun). Therefore, in this experiment, we fixed the hyperparameter of the central term to $k = 3$, while setting the value range of k_1 and k_2 between 1 and 9.

4.3 Results and Discussion

We use four common evaluation metrics, i.e., accuracy, precision, recall, and F1 score, to evaluate our approach.

Experiment 1. The results of Experiment 1 are detailed in Table 2. It is clear that the method we designed achieves the best level of performance. On the TroFi and MOH-X datasets, our method improves 40.8% and 37.3%, respectively, compared to predicting F1 values directly using ChatGPT. This suggests that by bootstrapping and combining GPT-generated surface knowledge, such as common literal collocations of verbs, to the domain of metaphor detection, it is possible to significantly improve GPT's performance in detecting verb metaphors. In addition, compared to the unsupervised strong baseline (Concrete-Abstract), the performance of our method improves by 8.3% and 29%, further demonstrating the superiority of our designed method on the unsupervised verb metaphor detection task.

Experiment 2. As shown in Table 3, the best performance on the entire TroFi dataset was obtained using the WordNet theme, which achieved an F1 score of 61.0%. While on the MOX dataset, the best performance was obtained using the Oxford Dictionary theme with an F1 score of 70.1%. For the hyperparameter k, we observed no significant performance difference between the two datasets by setting k to 1 or 3 when using WordNet topics or LDA topics. However, setting k to 3 slightly improves the performance when using the Oxford Dictionary theme. This may be due to the presence of polysemy in Oxford topics, i.e., different noun meanings correspond to multiple topic information, thus extending the scope of the verb table to cover literal topics. In addition, the performance of the three topic types is relatively close in the test results on the TroFi dataset, whereas on the MOX dataset, the WordNet topic and the LDA topic perform sim-

Models	TroFi				MOX-H			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
WordNet_Topic	46.0	96.8	44.6	**61.0**	53.6	90.1	51.4	65.4
WordNet_Topic_k	46.2	95.9	44.5	60.6	54.1	88.6	51.7	65.3
LDA_Topic	45.9	91.4	44.2	59.6	51.2	94.0	50.0	65.3
LDA_Topic_k	44.5	96.9	43.9	60.4	52.2	92.9	50.3	65.3
Oxford_Topic	47.0	90.4	44.6	59.8	62.9	86.7	58.1	69.6
Oxford_Topic_k	45.8	93.7	44.2	60.1	61.2	93.3	56.1	**70.1**

Table 3: Performance comparison on MOH-X and TroFi datasets using different topic mappings. The WordNet_Topic, LDA_Topic, and Oxford_Topic represent three different topics, respectively. The ones ending with "k" indicate that the first three nouns are extracted as the center nouns, while the ones without "k" indicate that one is extracted.

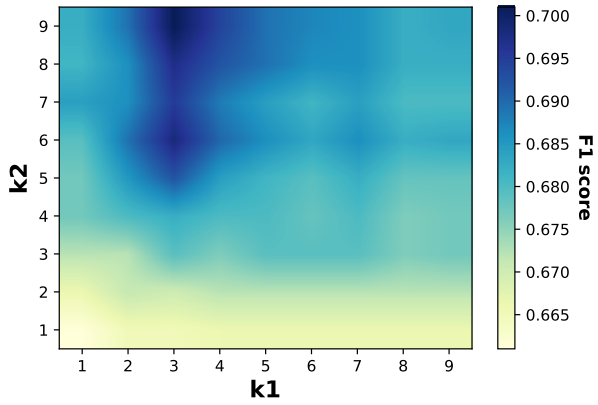

Figure 1: Effect of parameters $k1$, $k2$ on model performance, where k_1 represents the number of literal or non-metaphorical collocations selected from the verb list and k_2 denotes the number of topics that may be covered by the subject and object corresponding to the target verb.

ilarly, whereas the F1 score of the Oxford topic is higher than that of the other two topics (4.8%).

Experiment 3. Detailed results are presented in Fig 1. We used only the MOH-X dataset and kept the hyperparameter k at a fixed value of 3. This performance improvement can be attributed to the fact that increasing k_1 introduces more literal collocations from the verb list. As a result, this makes the model more capable of detecting the non-metaphorical content associated with a particular verb, thus reducing the number of misclassifications.

In addition, performance peaks when the hyperparameter k_2 is set to 3. However, when continuing to increase the value of k_2, the model's performance in detecting metaphors decreases instead. This suggests that considering multiple meanings of the focal word may introduce metaphorical information or redundant topics, which may affect the performance. Thus, our experimental results emphasize the need to weigh the model performance and the impact of theme introduction when choosing the value of k_2.

5 Conclusion

We present a novel approach aimed at improving the performance of an unsupervised verb metaphor detection task using model knowledge from ChatGPT. This approach does not rely on hand-coded knowledge or manually labeled datasets. First, we construct a literal meaning collocation lookup table for each target verb. When parsing the input text, we pay special attention to the subjects and objects corresponding to the verbs to be detected. We introduced a variety of topics, including WordNet topics, LDA topics, and Oxford topics. By comparing the relationship between subject and object topics in the input text and the verb topics in the verb table, we determine whether the text contains metaphorical expressions. The results show that by delicately combining and directing the model knowledge, we are able to significantly improve the performance of ChatGPT in the verb metaphor detection task.

6 Limitations

We introduce a verb table containing literal subject-verb and verb-object collocations for each target vocabulary. However, the literal collocations generated using ChatGPT are not always comprehensive, which leads to some literal samples being incorrectly categorized as metaphorical usage. In addition, due to varying syntactic structures, when ana-

lyzing subject-verb-object relations in input texts using ChatGPT, there may be parsing errors or structures that are not present, which also affects the performance of the overall method. In future work, we would like to investigate more powerful generative models or natural language parsing tools to improve the coverage of literal collocations in verb lists or to improve the accuracy of parsing subject-verb-object relations of input texts.

7 Ethics Statement

Metaphor, as a linguistic phenomenon that conveys implicit semantics, is capable of concretizing abstract concepts or enriching substantive concepts. This makes it possible for metaphors to be used as a tool for communicating political positions and gaining voter support in the political domain. However, our proposed zero-shot metaphor detection approach can also be used to identify metaphorical expressions and address the above issues from a governance perspective. In addition, we advocate the inclusion of tasks related to metaphor detection and generation, especially the application of ChatGPT to downstream metaphor applications, into the AI ethical code.

References

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.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

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

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

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. *arXiv preprint arXiv:2104.13615*.

Lisa Gandy, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 328–334.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66.

Mary Hesse. 1965. Models and analogies in science.

Yael Karov and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational linguistics*, 24(1):41–59.

Adam Kilgarriff. 2000. Wordnet: An electronic lexical database.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20.

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

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13534–13542.

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*. Association for Computational Linguistics (ACL).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Malay Pramanick and Pabitra Mitra. 2018. Unsupervised detection of metaphorical adjective-noun pairs. In *Proceedings of the Workshop on Figurative Language Processing*, pages 76–80.

KC Shet, U Dinesh Acharya, et al. 2012. A new similarity measure for taxonomy based on edge counting. *arXiv preprint arXiv:1211.4709*.

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*, pages 160–170.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 978–988.

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012: Posters*, pages 1121–1130.

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*, pages 30–39.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

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.

Shun Wang, Yucheng Li, Chenghua Lin, Loïc Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising. *arXiv preprint arXiv:2302.05611*.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44.

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

8 Appendix

The main purpose of this section is to detail how GPT-3 can be utilized to obtain literal collocations of verbs, as well as to obtain the required cues for subject and object pairs in the input text.

8.1 Analyzing Literal Collocations

For verb literal collocation parsing, we assume that the target verb is w_k. We do this by making a request to GPT-3 to generate all possible literal collocations of w_k, including both subject-predicate and predicate-object parts. We explicitly labeled the desired output format at the end of the request:

> Please provide as many subject and object topic categories as possible that are paired with the verb '$\omega_{\mathbf{k}}$' in non metaphorical or literal usage. The format is: Subject Categories:
> 1.
> 2.
> Object Categories:
> 1.
> 2.

8.2 Analyze Subject-Object Pairs

For subject-object parsing of the input text, we consider a specific target verb w_k, whose corresponding context is S, and the position of the verb w_k in the context is indicated by the index k. We make a request to GPT-3 to generate the subject and object corresponding to the verb w_k in the context. Again, we explicitly labeled the desired output format at the end of the request:

> For the sentence '$\mathbf{S}$'. Give the subject and object of the verb '$\omega_{\mathbf{k}}$' located in '$\mathbf{k}$' in order of format. For example,
> subject:
> object: