Filling the Last Gap: Introducing Multi-Word Expressions to Verb Metaphor Detection

Anonymous ACL submission

Abstract

Metaphor, as a powerful cognitive modality, possesses the ability to transfer knowledge structures from one domain to another. As metaphor detection continues to receive attention in the field of natural language processing, its importance in downstream tasks such as information extraction, sentiment analysis, and human-computer interaction has gradually become more prominent. However, previous studies have mainly focused on the implicit semantics of individual words, ignoring the fact that combinatorial words may have implicit semantics. In this paper, we propose for the first time a verb metaphor detection task containing multiple words. The goal of this task is to identify verbs or verb phrases with metaphorical usage in a sentence. Subsequently, we introduced a new dataset of verb metaphors. Next, we employed the theory of selection preference violation (SPV) and the metaphor identification program (MIP) for the multi-word verb metaphor task, both of which have been shown to be effective in single-verb metaphor detection. The experimental results show that SPV and MIP can effectively improve the performance of the model on the multi-word verb metaphor detection task.

1 Introduction

Metaphor is a rhetorical device in metaphorical language (Abulaish et al., 2020) that uses specific words to represent another concept in a given context (Krishnakumaran and Zhu, 2007), thus conveying an analogy between two seemingly unrelated concepts (Fass, 1991). As metaphor research continues, metaphor detection has shown potential to improve the accuracy of downstream natural language processing (NLP) tasks (Veale et al., 2015), including sentiment analysis and text categorization. In addition, it can even enhance a model's ability to understand multimodal image information (Akula et al., 2022).

In the task of Verb Metaphor (VM) detection, previous studies have typically used Selection Preference Violation (SPV) (Wilks, 1975, 1978) and Metaphor Identification Program (MIP) (Group, 2007) for metaphor identification. SPV describes the metaphorical phenomenon that occurs when selective preferences in the context of a verb are broken. For example, in the sentence "The flowers whispered to each other.", the verb "whispered" with a non-human collocation (i.e., "flowers" is a non-preferred word) constitutes a case of selective preference violation. MIP, on the other hand, judges metaphors on the basis of whether the underlying meaning of the target verb is consistent with the meaning that the verb acquires in context. In the sentence "His spirits began to sink as he realized the challenges ahead.", for example, the verb "sink" in its base meaning is "to dive into the water", whereas the meaning in the context is "to be depressed".

Although SPV and MIP have achieved good performance gains in metaphor detection, both methods focus on a single target verb and ignore the case of Verb Multi-Word Expressions (VMWE). Consider an example of VMWE:

The plane **took off** from the runway.

In this example, solely considering the individual meanings of the verb "take," such as "to physically pick up" or "to accept or receive something offered or given," does not align well with the noun "plane." However, when we consider "take off" as a holistic expression, encompassing meanings like "the action of removing or disrobing" or "the moment when an aircraft leaves the ground and begins its ascent into the air," it becomes evident that within the context, the association of "plane" with the second meaning is coherent. This clearly demonstrates that understanding the full range of metaphorical expressions necessitates consideration of the multi-word contextual usage of verbs rather than a singular interpretation of individual verbs.

Verb Multi-Word Expressions (VMWE) can be defined as "special interpretations that cross the boundaries of a single verb" (Sag et al., 2002), and the main focus of this definition is on the mismatch between the overall interpretation of a VMWE and the standard meanings of the individual words that make up the expression. To recognize VMWE, researchers need to consider the lexical combination as a whole (Calzolari et al., 2002) and make judgments in context, which is similar to the principle of Verb Metaphor (VM) detection (Wilks, 1978; Group, 2007).

Inspired by VM and VMWE, we introduce a new task, the Multi-word Verb Metaphor Detection (MVMD) task. The goal of this task is to determine whether a verb or verb combination uses metaphorical usage in a given context. Specifically, the contributions of this paper are as follows:

1. We are the first to introduce the Multi-word Verb Metaphor Detection (MVMD) task, where verb metaphors include both single-verb metaphors and combined verb metaphors.

2. We propose a multi-word verb metaphor dataset, which is a combination of the current mainstream verb metaphor dataset and verb multi-word metaphor dataset.

3. We apply the theory of Selection Preference Violation (SPV) and the Metaphor Identification Program (MIP) to the task of MVMD. The experimental results show that by directing the model to focus on verb combinations, the performance of the model on the MVMD task can be effectively improved.

2 Preliminaries

In this section, we will provide a brief introduction to the concepts of multi-word expressions and metaphors. In §2.1, we will delve into theories related to metaphor. Subsequently, in §2.2, we will introduce the related aspects of multi-word expressions and verb multi-word expressions, respectively.

2.1 The theory of metaphors

Metaphors are a rhetorical device in metaphorical language (Abulaish et al., 2020). They refer to entities that are similar to the objects to which they refer in a literal interpretation (Egg and Kordoni, 2023). Metaphors represent another concept by using one or more words in a given context rather than adopting the literal meaning of the expression (Fass, 1991). Lakoff and Johnson (1980) proposed Conceptual Metaphor Theory (CMT). CMT categorizes metaphor as a conceptual mapping between source and target domains and gives the definition "In metaphor, there are two domains: the target domain, which consists of the immediate subject matter, and the source domain, where significant metaphorical reasoning occurs and provides the source concepts used in the reasoning". For example, "Life is a journey". By reasoning metaphorically between the source domain (life) and the target domain (journey), an implicit meaning or point of view about life is conveyed. Wilks (1975, 1978) developed Selective Preference Violation (SPV). They argue that metaphors occur when selective preferences in context are broken. However, not all preference violations constitute metaphors (Ge et al., 2023). For example, traditional metaphors evolve into literal meanings as people use them frequently.

2.2 Multi-word Expression

Multi-Word Expression. Multi-Word Expression (MWE) are an important object of study in natural language processing. Villavicencio et al. (2005b) emphasized that identifying MWE is crucial to ensure that the system maintains meaning, generates appropriate translations, and avoids producing unnatural or meaningless sentences. However, there are some differences in the conceptualization of MWE among different research scholars. Sag et al. (2002) defines MWE as "special interpretations that cross word boundaries (or spaces)", emphasizing that the overall meaning of MWE does not match the standard meanings of the individual words that make up the expression. MWE include fixed expressions, semi-fixed expressions, and syntactically flexible expressions. Further, MWE include idioms, compound nouns, proper names, verb-particle constructions, institutionalized phrases, and light verbs. A more general definition is provided by Calzolari et al. (2002), which considers MWE as "sequences of words that act as individual units at some level of linguistic analysis", characterized by high lexicalization, reduced combinativity, and rule violations. Alegria et al. (2004), on the other hand, treats multi-word expressions as including a variety of word combinations, ranging from idioms, proper names, compound words, lexical and grammatical collocations to institutionalized phrases.

Verb Multi-Word Expression. Verb Multi-Word Expressions (VMWE) is a particularly challenging subcategory of MWE (Waszczuk et al., 2019). VMWE consists mainly of Light Verb Constructions (LVC), Verb-Particle Constructions (VPC), and idioms. Among these, LVC is a combination of a verb and a noun, where the verb loses its meaning to some extent while the noun retains one of its original meanings (Sag et al., 2002), e.g., "take a walk". VPC consists of a verb and one or more particles (Sag et al., 2002), e.g., "brush up on". An idiom is a phrase (or sentence) that is habitually used with a meaning different from the literal meaning of its construction (Villavicencio et al., 2005b). (Sag et al., 2002) categorizes idioms into two types, an indecomposable class that is not affected by syntactic changes because it is semantically opaque, e.g., "bite the dust". The other is a decomposable category with varying degrees of syntactic variation, which is more grammatically flexible, e.g. "open a can of worms". Since the detection of idioms does not depend on context (Villavicencio et al., 2005a), this conflicts with the definition of verb implicit semantics that we introduced. Therefore, we do not take the idiomatic part of VMWE into account in our study.

VMWE have attracted the attention of researchers as a particularly challenging subclass of Multi-Word Expressions (MWE) due to their properties such as incoherence, overlap, different word order, and syntactic or semantic ambiguity (Waszczuk et al., 2019). Since the detection of idioms is not context-dependent (Villavicencio et al., 2005b), this conflicts with the definition of verb implicit semantics that we introduced. Therefore, we do not take the idiomatic part of VMWE into account in our study.

3 Related Work

3.1 Supervised Metaphor Detection

Currently, metaphor detection tasks are mainly focused on supervised methods. For example, Mao et al. (2019) employed generic corpus information as context to detect metaphors using MIP and SPV paradigms. Le et al. (2020), on the other hand, attempted to apply dependency tree knowledge to metaphor detection by constructing graph network adjacency matrices in order to utilize dependency tree structure information. For knowledge injection, Li et al. (2023b) used two encoders, one of which was fine-tuned by FrameNet (Fillmore et al., 2002). Choi et al. (2021) applied MIP and SPV to pre-trained models. To improve the detection performance of BERT, Zhang and Liu (2022); Li et al. (2023a) introduced example sentences as a control. While Zhang and Liu (2022) used literal meaning samples from the original dataset, Li et al. (2023a) introduced example sentences from a dictionary. Su et al. (2021); Babieno et al. (2022) introduced the underlying meaning of the target word directly. More recently, Badathala et al. (2023); Zhang and Liu (2023) attempted to introduce multi-task learning. Badathala et al. (2023) introduced exaggerated corpus knowledge into metaphor detection, while Zhang and Liu (2023) introduced a word sense disambiguation task and used adversarial learning (Ganin and Lempitsky, 2015) to guide the model to learn the data distributions for both tasks, achieving the best performance in the metaphor detection task so far.

3.2 Multi-word Expression Detection

Currently, common approaches for recognizing MWE include rule-based systems (Foufi et al., 2017; Pasquer et al., 2020), Conditional Random Fields (CRF)-based systems (Liu et al., 2020; Kishorjit et al., 2011), and labeled word-level systems (Rohanian et al., 2019; Savary et al., 2019). Among these approaches, rule-based systems remain competitive with neural models, while many also use MWE dictionaries to aid in MWE detection (Tanner and Hoffman, 2023). Some approaches, e.g., (Tanner and Hoffman, 2023), employ a similar approach to Word Sense Disambiguation (WSD) using dual encoders, introducing an innovative multi-encoder architecture that addresses both MWE detection and WSD. Another related work (Kanclerz and Piasecki, 2022) uses a similar approach to (Tanner and Hoffman, 2023) to model the MWE detection task as a classification problem.

4 Method

4.1 Mission Description

Multi-word Verb Metaphor Detection task. In previous studies, verb metaphors refer to the meaning of a verb conveyed in a particular context, which is usually not a direct extension of its literal meaning. For example, metaphor detection systems (Choi et al., 2021; Zhang and Liu, 2022; Li et al., 2023a) employ the theory of selection preference violation (MIP) (Group, 2007) to determine the presence of metaphors by comparing the underlying meaning of the target word with the meaning of the context. This has similarities with the Verb Multi-Word Expression (VMWE) detection task. In VMWE, the overall semantics is independent of the individual segments and the overall collocation cannot be replaced by synonyms (Constant et al., 2017). In addition, the meanings of verbs in context are often considered non-literal; they are usually treated as non-literal except for idioms (which usually have agreed-upon literal meanings). Therefore, inspired by the above phenomenon, we merged the Verb Metaphor (VM) and Verb Multi-word Expression (VMWE) tasks into the Multi-Word Verb Anaphora Detection task. The goal of this task is to help the model understand and recognize combinatorial verbs simultaneously while recognizing verb metaphors.

Data Labeling Methods. According to the literature (Constant et al., 2017), the Multi-Word Expression task consists of two main parts: discovery and detection. The former is usually used to find new MWE types in a text corpus, while the latter involves automatically annotating MWE in text using known MWE types. In MWE research, most of the literature (Walsh et al., 2022; Schneider et al., 2016; Swaminathan and Cook, 2023; Premasiri and Ranasinghe, 2022) adopts token-level based annotation methods, and some studies directly output VMWE types (Yirmibeşoğlu and Güngör, 2020) (e.g., VID) or directly annotate whether they are VMWE (Boukobza and Rappoport, 2009). The VMWE set by Yirmibeşoğlu and Güngör (2020); Boukobza and Rappoport (2009) do not take context into account, but give direct multi-word combinations, e.g. (Verb, Preposition, Noun). This is in some conflict with our defined task, which requires context-based metaphorical inference. For this reason, this paper employs token-level annotation to annotate the dataset. token-level annotation aims to categorize each token (usually words or subwords) in a text by assigning a label or category to each token.

In VMWE annotation, some studies (Zaninello and Birch, 2020; Vincze et al., 2011) have used the Inside-Outside-Beginning (IOB) annotation approach. In the IOB annotation approach, each element (usually words or tokens) is labeled as B

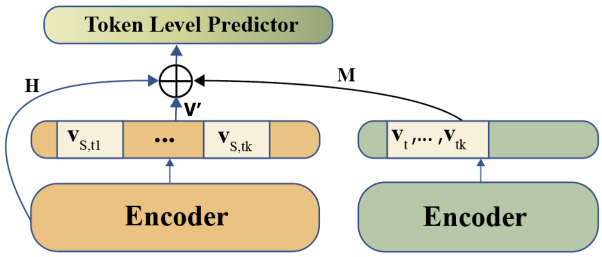

Figure 1: Model structure diagram. H is the full contextual features, $V' = \mathbf{v}_{S,ti},\ 1 \leq i \leq k$ is the contextual features of the k constituents of the verb phrase. $M = \mathbf{v}_{ti},\ 1 \leq i \leq k$, for the underlying meaning of the verb or verb phrase. The result of the integration of the three features will be used for token-level classification prediction.

(entity beginnings), I (internal parts of entities), or O (outside entities). However, VMWE may have discontinuous parts. To solve this problem, Schneider et al. (2016); Dyer and Smith extended the IOB labeling approach to eight tokens, which are "BbIiOo_ ". In contrast, the PARSEME dataset (Savary et al., 2023) uses the "VMWE type" and "*" annotation. Consider the following example:

Great , we look forward to seeing you
* * * 1:VPC.full;2:IAV 1;2 2 * * *.

In this case, the target verb "look" is labeled with two VMWE categories, VPC.full and IAV, which are split by ";", while the combinatorial word "forward" corresponding to the verb is only labeled with 1 and 2, indicating the continuation of VUC.full and IAV. With this example, we can see that "look forward" is labeled as a multi-word expression of class VPC.full, while "look forward to" is labeled as a multi-word expression of class IAV. In order to adapt to the sequence annotation task, we simplify the annotation of PARSEME (Savary et al., 2023) to 0/1 annotation. That is, when a target verb or verb phrase is identified as having metaphorical usage, it is labeled as 1; conversely, it is labeled as 0. Specifically, verbs or verb phrases containing metaphorical expressions are labeled as 1, while the rest of the context or content outside the verb table is labeled as 0. We will describe the construction method of the verb table in detail in the dataset construction section of §5.2.

4.2 Model Design

The specific structure of the model designed in this paper is detailed in Fig. 1. MelBERT (Choi et al., 2021) was the first study to combine SPV and MIP into a pre-trained model and achieved good performance on a single-verb metaphor task. We extend the SPV and MIP methods to the verb polysemy domain. For SPV, we first use Boolean lists to extract the verb part of the hidden layer output V:

$$V' = [0, ...\mathbf{v}_{S,[t_1,...,t'_1]}, ..., \mathbf{v}_{S,[t_k,...,t'_k]}, ..., 0],$$

where $\mathbf{v}_{S,[t_i,...,t'_i]}$, $1 \leq i \leq k$ are the contextual features of the k constituents of the verb phrase, which are not necessarily continuous. For MIP, we use an Encoder (e.g., RoBERTa(Liu et al., 2019)) to extract the basic meaning of the verb with:

$$M = \mathbf{v}_{cls}, \mathbf{v}_{[t_1,...,t'_1]}, ...\mathbf{v}_{[t_k,...,t'_k]}, \mathbf{v}_{sep} = f_b([\mathrm{CLS}], w_{[t_1,...,t'_1]}, ..., w_{[t_k,...,t'_k]}, [\mathrm{SEP}]), \quad (1)$$

where $\mathbf{v}_{[t_i,...,t'_i]}$, $1 \leq i \leq k$ is the literal meaning of the verb group. Then, we combine the verb contextual meaning V', the verb basic meaning M and the whole context H proportionally, i.e:

$$H' = H + w_1 * V' + w_2 * M, \quad (2)$$

where H' is the final output, w_1, w_2 are the weight parameters for the SPV and MIP, respectively.

5 Dataset

This section describes in detail our multi-word verb metaphor dataset PVTM (PARSEME-VUA-Trofi-MOH). In §5.1, we discuss in detail the dataset required to construct PVTM. And in §5.2, we provide a detailed description of the preprocessing, construction, and segmentation approach of PVTM.

Dataset	Tokens	Sentences	% Met.
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	1,639	1,639	25.0%
TroFi	3,737	3,737	43.5%

Table 1: Dataset statistics. tr: training set. val: validation set. te: test set. tokens: number of vocabulary units or samples to be tested. sent.: total number of sentences, %Met.: metaphorical samples as a proportion of the total samples

5.1 Dataset Introduction

We introduced two types of datasets covering verb metaphors and verb multi-word expressions, respectively. Specifically, the verb metaphor dataset includes VUAverb, TroFi, and MOH-X, while the verb multi-word expression dataset is PARSEME.

TroFi. The TroFi dataset (Birke and Sarkar, 2006) is derived from the Wall Street Journal corpus (Charniak et al., 2000). In the original TroFi dataset, each sample is annotated with one of three labels: l (literal), n (non-literal), or u (unannotated). We used the (Choi et al., 2021; Zhang and Liu, 2023) version of the TroFi dataset, which includes literal and metaphorical usage of 50 English verbs, totaling 3,717 samples, as examples of verb metaphors.

MOH. The MOH dataset was originally created by Mohammad et al. (2016), and its construction methodology involves first extracting polysemous verb samples from WordNet, and then metaphorically labeling the sentences via a crowdsourcing platform. To ensure the quality of the dataset annotation, Mohammad et al. (2016) adopted a 70% annotation consistency criterion. A subset of MOH, MOH-X (Shutova et al., 2016), which references mainstream metaphor detection systems (Choi et al., 2021; Zhang and Liu, 2023), excludes instances with pronouns, subordinate subjects or objects. In this paper, we consider the full MOH data.

VUAverb. The VU Amsterdam Metaphor Corpus (Steen et al., 2010) [1] metaphorically annotates each lexical unit in a subset of the British National Corpus (BNC) (Edition et al.). The annotation was done using the MIPVU program, with high inter-annotator agreement and Kappa values greater than 0.8. Based on VUAMC, several different variants of the VUA corpus have emerged, among which VUAverb is the verb version of the VUA corpus. In this paper, we use the dataset mentioned in the metaphor detection shared task (Leong et al., 2018, 2020). We merged the training, validation and test sets of VUAverb, which included a total of 22,668 samples.

PARSEME. PARSEME is a multilingual MWE corpus, developed by an international community, and is one of the most widely used datasets in VMWE research.The annotation of PARSEME was performed using a method based on the XML (van Gompel and Reynaert, 2013) annotation format, via a Web platform. The English section was first introduced in version 1.1 (Walsh et al., 2018), and the data sources include the English-EWT corpus (Silveira et al., 2014), the LinES parallel corpus

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

Dataset	Sent.	VMWE	LVC.full	LVC.cause	VPC.full	VPC.semi	IAV	MVC	VID
Train	1878	271	97	12	112	16	22	12	44
Dev	1132	169	63	10	62	7	13	9	35
Test	3466	517	172	29	194	30	36	29	108
Total	6476	957	332	51	368	53	71	50	187

Table 2: PARSEME dataset statistics. sent.: total number of sentences. VMWE: number of verb VMWE. LVC: Light-Verb Constructions, including both LVC.full and LVC.cause. VPC: Verb-Particle Constructions, including VPC.full and VPC.semi. IAV: Inherently Adpositional Verbs. MVC: Multi-Variable Construction. VID: Verbal Idiom.

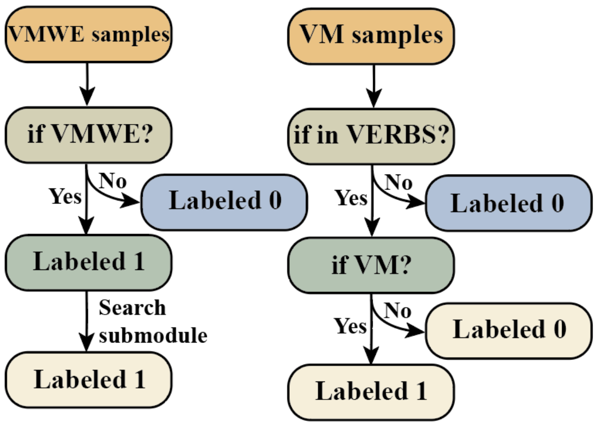

Figure 2: Flowchart of the dataset labeling process. For multi-word samples, if the target verb is a VMWE, both the verb and its combinations are labeled as 1. For verb metaphor (VM) samples, verbs not in VERBs are labeled as 0 (VERBs denote the set of verbs occurring in PARSEME). For verbs within VERBs, metaphorical usage was labeled as 1 and non-metaphorical usage was labeled as 0.

(Ahrenberg, 2007), and the Parallel Universal Dependencies (PUD) treebank (Zeman et al., 2018). In this study, we chose PARSEME 1.3 (Savary et al., 2023), which contains the VMWE portion of the PVTM dataset. version 1.3 of the English corpus has been pre-parsed. Similar to the metaphor dataset, we merged the partitioned dataset, which included a total of 6476 samples.

5.2 Dataset Construction

Combination of Dataset. In the token-level annotation task, our goal is to identify whether there are implicit semantic expressions in the context that are related to the verb set, which may include one or more verbs and the VMWE collocations associated with those verbs.

For the PARSEME dataset, we used two main steps. First, we merged the samples labeled as VMWE directly into PVTM and labeled such samples as non-literal. Second, based on the set of verbs tagged as VMWE in the PARSEME dataset (VERBS), we expanded the samples that were not tagged as VMWE. In these samples, we first label the verbs present in the VERBS, and then identify the combinations of verbs that correspond to these verbs in a sentence (if present) and label them with their literal meanings. Eventually, these samples will be merged into PVTM. For the metaphor dataset, we merge samples from the VUAverb, TroFi, and MOH-X datasets that contain VERBS verbs into PVTM. Specifically, we merged the samples from the VUAverb dataset by combining different verb samples from the same sentence into a single record. Since the same sentence in the TroFi and MOH datasets does not contain more than one verb to be detected, there is no need to merge the samples from TroFi and MOH.

Dataset labeling. The PVTM dataset labeling process is illustrated in Fig 2. PVTM considered only verbs that appeared in PARSEME for the VMWE samples, called VERBs. for the multi-word samples, the VMWE were labeled as 1, and the remaining contexts as 0. For the verb metaphor samples, verbs of metaphorical usage that existed within VERBs were labeled as 1, and verbs that did not exist within VERBs, or VERBs within verbs with non-metaphorical usage are labeled as 0.

Dataset segmentation. To ensure that the partitioned datasets have similar data distributions, we considered four key aspects of PVTM for partitioning: verbs, verb types (literal meaning, metaphorical or multi-word), labels (literal or non-literal), and dataset types (PARSEME, VUA, TroFi, and MOH). We divided the whole dataset into training, development and test sets with a division ratio of 0.7, 0.15, 0.15. For the cases where some categories contain only one or two samples, we similarly assigned to one of the three subsets according to the above ratio. In PVTM, the training set contains 4474 samples, the development set contains 1066 samples, and the test set contains 1053 samples.

6 Experiments

This section evaluates the performance of the baseline model on the TVPM dataset. In §6.1, we provide an introduction to the traditional baseline model. And in §6.2 and §6.3, we present the content of the experiments and the parameter details of the experimental execution, respectively.

6.1 Baseline Model

BERT (Devlin et al., 2018) is a bi-directional coding model based on the Transformer architecture, proposed in 2019. The model employs two self-supervised learning strategies. One of them is the Masked Language Model (MLM) strategy which aims to randomly mask a certain percentage of input tokens and then let the model predict these masked tokens. The other strategy, Next Sentence Prediction (NSP), is used to predict the coherence between sentences. For example, given two sentences A and B, the model will mark them as "IsNext" if they are contextual; if B is randomly selected from other sentences, the model will mark them as "NotNext". RoBERTa (Liu et al., 2019), on the other hand, improves on BERT (Devlin et al., 2018) by employing a more domain-specific English corpus for training. Its self-supervised training strategy is similar to that of BERT, which includes MLM and NSP.In this experiment, we use BERT and RoBERTa as baseline models. For each model type, we only considered the BASE version.

Models	**Token Level**		
	Pre.	**Rec.**	**F1**
BERT$_{bs}$	24.0%	40.0%	30.0%
RoBERTa$_{bs}$	29.4%	32.1%	30.7%
RoBERTa$_{bs} + s$	38.9%	39.9%	39.4%
RoBERTa$_{bs} + m$	37.1%	42.6%	39.7%
RoBERTa$_{bs} + sm$	37.0%	46.3%	41.1%

Table 3: Model evaluation results. BERT$_{bs}$: BERT-base. RoBERTa$_{bs}$: RoBERTa-base. s: Selection preference violation (SPV). m: Metaphor Identification Program (MIP). sm: SPV and MIP.

6.2 Experimental Design

The token-level annotation task requires the classification of the hidden layer output of an entire sentence. In comparing the two baseline models, BERT and RoBERTa, we chose to use BERT-base and RoBERTa-base as control models. In addition, we introduced three additional baseline models, RoBERTa-base+SPV, RoBERTa-base+MIP, and RoBERTa-base+SPV+MIP. these are denoted in the experimental results as RoBERTa$_{bs}$+s, RoBERTa$_{bs}$+m, and RoBERTa $_{bs}$+sm.

In the model designed in 4.2, Eq 2 contains two hyperparameters, w_1 and w_2, which are used to control the extent of combining SPV and MIP information. In this experiment, we choose RoBERTa-base to conduct experiments on the PVTM dataset with the aim of exploring the effect of these two hyperparameters on the F1 performance of the model. The search range of w_1 and w_2 is set from 0.1 to 1.5 with an interval of 0.1. We designed three sets of experiments, namely, single w_1, single w_2, and the combination of considering w_1 and w_2.

6.3 Implementation

In this experiment, we use a similar experimental setup as in (Choi et al., 2021). We used the Adam (Kingma and Ba, 2014) optimizer with an initialized learning rate of 3e-5; the learning rate was controlled by a linear warmup scheduler, and the learning rate was gradually increased during the warmup period, with warmup epoch set to 3. We set a dropout rate of 0.2. The size of the hidden layer was set to 768. the batch sizes for both training and validation, and testing were set to 100, and the maximum number of training rounds was set to 15. the maximum length of sentences was limited to 150 tokens. we set the weights to 150 to balance out the lower percentage of verb-metaphor content in the sample. All experiments were run on a cloud server equipped with a single card A100 80G GPU.

7 Evaluation of Metric and Results

7.1 Evaluation Metric

For metaphor detection tasks, previous studies (Choi et al., 2021; Zhang and Liu, 2022; Li et al., 2023a) 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 are truly

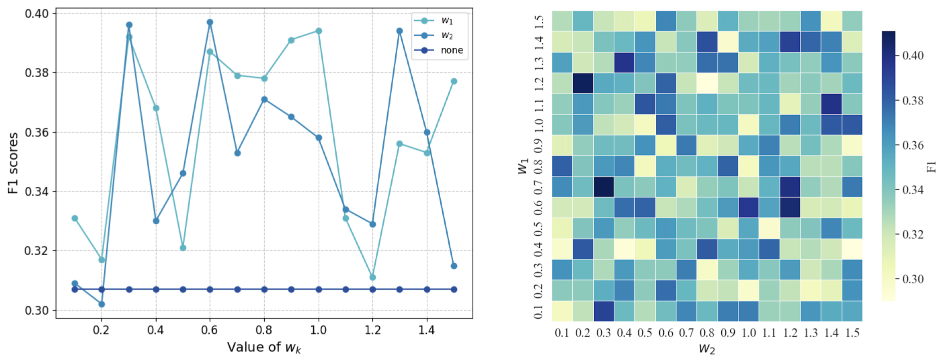

Figure 3: The hyperparameter analysis plots are shown below. The line graph on the left side presents the effect of using a single w_1, w_2 on the F1 performance of the roberta-base model. The heatmap on the right side presents the effect of using a combination of w_1 and w_2 on the model F1 performance.

positive categories among those determined by the model to be positive categories, 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 with its Recall. Multi-word expression detection is similar to the metaphor detection task, and previous studies (Ramisch et al., 2023; Swaminathan and Cook, 2023) mainly used the F1 score as the main evaluation metric, while Sarlak et al. (2023); Savary et al. (2023) considered precision, recall and F1 score together. In this experiment, we considered accuracy, precision, recall and F1 score simultaneously.

7.2 Analysis of results

The experimental results are presented in Table 3. The study shows that the independent use of SPV, MIP or their combination significantly improves the model's performance on token-level tasks (8.7%, 9.0%, and 10.4% higher, respectively). Particularly noteworthy is that the model with a specific combination of SPV and MIP reached the highest level of F1 value at 41.1%. This suggests that the SPV and MIP structure can correctly guide the model to focus on the difference between the contextual and literal meanings of verbs or phrases, thus improving the model's performance on token-level annotation tasks.

In Fig.3 left, we investigate the effect of using a single w_1 (SPV) and a single w_2 (MIP) on the performance of roberta-base F1 in token-level tasks. The results show that in most cases, the model performs better when SPV and MIP are used alone compared to when they are not. The model achieves the highest F1 when $w_1 = 1.0$ or $w_2 = 0.6$, 39.4% and 39.6%, respectively. Figure 3 right shows the effect of combining w_1, w_2 on the model F1 performance. As can be seen from the figure, the proportion of correct combinations is higher than the baseline (without SPV and MIP) and even higher than with a single SPV or MIP. the model reaches its highest performance (F1=41.1%) when $w_1 = 0.7, w_2 = 0.3$. However, incorrect combination ratios can even cause the model to fall below the baseline, e.g., $w_1 = 1.2, w_2 = 0.8$ or $w_1 = 0.4, w_2 = 1.5$, at which point the model's F1 is 29.0% (1.7% below the baseline).

8 Conclusion

This study focuses on the task of verb metaphor detection at different levels of granularity, considering traditional verb metaphors and focusing on multi-word expressions of verbs. We propose a multi-word verb metaphor dataset, PVTM. this dataset integrates three classical datasets in the field of metaphor detection (including VUAverb, TroFi, and MOH), as well as a shared corpus in the field of verb multi-word expressions, PARSEME. in PARSEME, we consider groups of verbs other than verbal idioms to be Verb Multi-Word Expressions (VMWE). the PVTM dataset was labeled with token-level annotation. Meanwhile, we chose BERT and RoBERTa as baseline models and introduced SPV and MIP structures. The experimental results show that compared with direct prediction, directing the model to focus on verbs and verb combinations can significantly improve the model's performance in the verb-multiple-word anaphora detection task.

9 Limitations

This study proposes a new task, namely multi-word verb anaphora detection, and integrates current classical datasets in the field of anaphora and multi-word expressions. For the PARSEME dataset, we did not include the idiomatic part, which may result in a dataset that fails to comprehensively cover the various types of verb implicit semantics, thus presenting some challenges in fine-tuning the model's generalization ability. In addition, since the anaphoric or multi-word expression datasets are manually labeled, there is inevitably some noise, and combining them may introduce more noise. Finally, the timeliness of the dataset may also be problematic because the implicit semantics of some verbs may gradually evolve into literal meanings as people use the language. This may result in some verbs that are currently considered to have literal meanings being incorrectly labeled as implicit semantic usage.

In future research, we plan to extend the scope of implicit semantics to consider not only verbs, but also to explore the implicit semantics of other linguistic elements. In addition, we will also deal with the noise and timeliness issues of the dataset more carefully to improve the performance and generalization ability of the model.

10 Ethics Statement

The datasets used and research papers cited in this study were derived from publicly available sources, and we strictly adhered to the guidelines of academic and research ethics. We emphasized transparency and openness of information by providing explicit citations to the cited public data sources in order to fully respect the original authors and data providers of research related to the field of metaphor detection. This is in line with the principle of academic integrity and ensures full detection of the work and contributions of those who have gone before us. We will continue to uphold this principle in order to promote openness and cooperation in academic research.

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 (TWEB)*, 14(1):1–52.

Lars Ahrenberg. 2007. Lines: An english-swedish parallel treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 270–273.

ArjunR. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, WilliamT. Freeman, Yuanzhen Li, and Varun Jampani. 2022. Metaclue: Towards comprehensive visual metaphors research.

Inaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and treatment of multiword expressions in basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55.

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, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. *arXiv preprint arXiv:2305.17480*.

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.

Ram Boukobza and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 468–477.

Nicoletta Calzolari, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *LREC*, volume 2, pages 1934–1940.

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*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nathan Schneider Emily Danchik Chris Dyer and Noah A Smith. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut.

B Edition, BNC Baby, and BNC Sampler. British national corpus.

Markus Egg and Valia Kordoni. 2023. A corpus of metaphors as register markers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 220–226.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational linguistics*, 17(1):49–90.

Charles J Fillmore, Collin F Baker, and Hiroaki Sato. 2002. The framenet database and software tools. In *LREC*.

Vasiliki Foufi, Luka Nerima, and Eric Wehrli. 2017. Parsing and mwe detection: Fips at the parseme shared task. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 54–59.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, pages 1–67.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Kamil Kanclerz and Maciej Piasecki. 2022. Deep neural representations for multiword expressions detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

N Kishorjit, L Dhiraj, N Bikramjit Singh, Ng Mayekleima Chanu, and B Sivaji. 2011. Identification of reduplicated multiword expressions using crf. CICLing.

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. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Duong Le, My Thai, and Thien Nguyen. 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*, volume 34, pages 8139–8146.

Chee Wee 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*, pages 18–29.

Chee Wee 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*, pages 56–66.

Yucheng Li, Shun Wang, Chenghua Lin, and Guerin Frank. 2023a. Metaphor detection via explicit basic meanings modelling. *arXiv preprint arXiv:2305.17268*.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loïc Barrault. 2023b. Framebert: Conceptual metaphor detection with frame embedding learning. *arXiv preprint arXiv:2302.04834*.

Nelson F Liu, Daniel Hershcovich, Michael Kranzlein, and Nathan Schneider. 2020. Lexical semantic recognition. *arXiv preprint arXiv:2004.15008*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*, pages 3888–3898.

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.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

Damith Premasiri and Tharindu Ranasinghe. 2022. Bert (s) to detect multiword expressions. *arXiv preprint arXiv:2208.07832*.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of mwe identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. *arXiv preprint arXiv:1902.10667*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Mahtab Sarlak, Yalda Yarandi, and Mehrnoush Shamsfard. 2023. Predicting compositionality of verbal multiword expressions in persian. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23.

Agata Savary, Silvio Ricardo Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, et al. 2023. Parseme corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559.

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.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In *LREC*, pages 2897–2904. Citeseer.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, et al. 2010. A method for linguistic metaphor identification. *Amsterdam: Benjamins*.

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*, pages 1280–1287.

Raghuraman Swaminathan and Paul Cook. 2023. Token-level identification of multiword expressions using pre-trained multilingual language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 1–6.

Joshua Tanner and Jacob Hoffman. 2023. Mwe as wsd: Solving multiword expression identification with word sense disambiguation. *arXiv preprint arXiv:2303.06623*.

Maarten van Gompel and Martin Reynaert. 2013. Folia: A practical xml format for linguistic annotation–a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2015. Applications of Metaphor Processing, page 109–117.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005a. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech amp; Language*, 19(4):365–377.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005b. Introduction to the special issue on multiword expressions: Having a crack at a hard nut.

Veronika Vincze, István Nagy, and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John Philip McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an annotated corpus of verbal mwes for english. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193–200.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2022. A bert's eye view: Identification of irish multiword expressions using pre-trained language models. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*, pages 89–99.

Jakub Waszczuk, Rafael Ehren, Regina Stodden, and Laura Kallmeyer. 2019. A neural graph-based approach to verbal mwe identification. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019)*, pages 114–124.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Yorick Wilks. 1978. Making preferences more active. *Artificial intelligence*, 11(3):197–223.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. Ermi at parseme shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159.

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