ContrastWSD: Enhancing Metaphor Detection with Word Sense Disambiguation Following the Metaphor Identification Procedure

Mohamad Elzohbi, Richard Zhao
University of Calgary
Calgary, Alberta, Canada T2N 1N4
{melzohbi,richard.zhao1}@ucalgary.ca

Abstract

This paper presents ContrastWSD, a RoBERTa-based metaphor detection model that integrates the Metaphor Identification Procedure (MIP) and Word Sense Disambiguation (WSD) to extract and contrast the contextual meaning with the basic meaning of a word to determine whether it is used metaphorically in a sentence. By utilizing the word senses derived from a WSD model, our model enhances the metaphor detection process and outperforms other methods that rely solely on contextual embeddings or integrate only the basic definitions and other external knowledge. We evaluate our approach on various benchmark datasets and compare it with strong baselines, indicating the effectiveness in advancing metaphor detection.

Keywords: Metaphor Detection, Word Sense Disambiguation, Metaphor Identification Procedure

1. Introduction

A metaphor, is a rhetorical device that compares, implicitly, two objects or concepts that are seemingly dissimilar but share symbolic or figurative similarities, with the intention of illuminating a fresh perspective and a more elaborate and nuanced comprehension of the world. Metaphors are not only intrinsic to creative writing, but they are also ubiquitous in human communication. Metaphors typically involve employing words in a manner that diverges from their basic definition, and their figurative sense is dependent on the context in which they are used. While novelty is an indicator of greater creativity in metaphors, sometimes they become widely used and established in the language, ultimately entering the lexicon as conventional metaphors, also known as dead metaphors (Lakoff and Johnson, 2008).

Automatic metaphor detection, the process of identifying metaphoric expressions within a given text, is essential for various Natural Language Processing (NLP) tasks, such as sentiment analysis, text paraphrasing, and machine translation (Mao et al., 2018). The development of metaphor detection models presents a significant challenge, as it requires the identification and analysis of both the basic and contextual meanings of words within their respective contexts, as recommended by the Metaphor Identification Procedure (MIP) (Praggle-jaz Group, 2007; Steen et al., 2010) (see Figure 1).

Early approaches relied on extensive manual efforts in feature engineering. However, with the advent of word embedding techniques and neural networks, more efficient and effective methods emerged for this task (Song et al., 2021). Notably, transformer-based models have demonstrated promising capabilities in detecting metaphors (Su et al., 2020; Choi et al., 2021; Babieno et al., 2022; Li et al., 2023). Despite these advancements, there is still scope for further improvement to simulate the Metaphor Identification Procedure effectively. Therefore, the main objective of this study is to investigate the efficacy of transformer-based models in word sense disambiguation and in following the systematic Metaphor Identification Procedure to extract and contrast between the contextual word sense and the basic definitions of a target word to enhance automatic metaphor detection.

Our proposed method is evaluated on established benchmark datasets, and the results demonstrate significant improvements. Comparatively, our model consistently achieves superior (and occasionally comparable) precision, recall, and F1 scores when compared to other recent and robust metaphor detection models.

2. Related Work

Metaphor detection has been a subject of active research in NLP for several years. Traditional approaches to metaphor detection have relied on hand-crafted or automatically acquired linguistic features, but recent advancements in NLP have resulted in the development of transformer-based pre-trained models that have demonstrated state-of-the-art performance across various NLP tasks (Choi et al., 2021; Song et al., 2021). DeepMet (Su et al., 2020) formulates metaphor detection as a reading comprehension task and uses a RoBERTa-based model that incorporates global and local text context, question information, and part-of-speech as features to detect metaphors.

MelBERT (Choi et al., 2021) is a RoBERTa-

LREC-COLING 2024, pages 3907–3915
20-25 May, 2024.

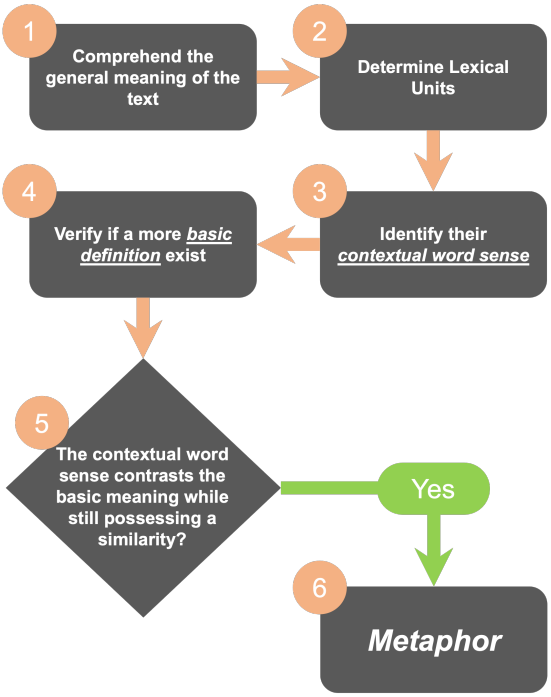

Figure 1: The Metaphor Identification Procedure

based model that incorporates linguistic metaphor identification theories. MelBERT captures the context-dependent nature of metaphors and has demonstrated state-of-the-art performance on multiple benchmark datasets. While the authors considered the MIP procedure in their design, their focus was on leveraging contextual and out-of-context word embeddings to represent the word sense and basic definitions of the word. However, utilizing contextual word embeddings may not always accurately represent the word sense definition; instead, it may lean more towards the general contextual meaning. Similarly, out-of-context word embeddings may not necessarily reflect the basic meaning of the word, as they may be influenced by the frequent meaning, which might not align with the word's basic sense (Pragglejaz Group, 2007; Babieno et al., 2022). In contrast, we encode both the contextual and basic definitions of the target words, which are extracted from the dictionary. This enables us to provide a more comprehensive understanding of their meanings and better align with the MIP procedure.

Researchers have also explored the use of external knowledge sources, such as definitions, word senses, and frames, to enhance the performance of metaphor detection models. For instance, Wan et al. (2021) used gloss definitions to improve metaphor detection by considering both contextual embedding and contextual definition that best fits the context. Similarly, Babieno et al. (2022) explored the integration of the most basic definitions from Wiktionary to improve MelBERT's performance, achieving comparable or superior results. In contrast, our model extracts the contextualized definitions and contrasts them with the basic definitions to align with the MIP procedure. FrameBERT (Li et al., 2023) proposed a new approach that incorporates FrameNet (Baker et al., 1998) embeddings to detect concept-level metaphors, achieving comparable or superior performance to existing models. Although encoding concepts may improve the model's understanding of similarities, it might not fully capture the variations in word meanings in various contexts.

Word-Sense Disambiguation (WSD), which involves identifying the correct sense of a word in context, is a challenging task in NLP with various applications. Our study shows that WSD can aid in the process of identifying metaphors by disambiguating the word sense in given contexts. Multiple state-of-the-art WSD models have been proposed, including a modified version of BERT (Yap et al., 2020) trained on a combination of gloss selection and example sentence classification objectives. Bevilacqua and Navigli (2020) propose a method for incorporating knowledge graph information into WSD systems to improve their performance. The authors use a large-scale knowledge graph (DBpedia) to provide additional context and semantic information for each word in the text. SenseBERT (Levine et al., 2020) pre-trains BERT on a large-scale sense-annotated corpus using a modified loss function to incorporate sense-aware training objectives.

Incorporating WSD models to obtain contextual definitions for use in metaphor detection has also been explored. For example, Metaphorical Polysemy Detection (MPD) (Maudslay and Teufel, 2022) has been proposed, which focuses on detecting conventional metaphors in WordNet. By combining MPD with a WSD model, this method can determine whether a target word represents a conventional metaphor within a given sentence. The authors have identified the two limitations mentioned earlier regarding MelBERT, which hinder its alignment with the MIP procedure. Particularly, attempting to implicitly infer word sense from the target word's contextual representation and assuming that the out-of-context embedding of the target word represents its basic meaning.

To address these issues, the authors trained an MPD model jointly with a WSD model to detect the metaphoricity of a target word, leveraging the word senses predicted by the WSD model for the target word in a given context. However, we argue that this model still lacks full alignment with MIP since it implicitly contrasts the basic definition with the word sense.

To achieve a more explicit alignment with MIP, we propose a different approach. Firstly, we utilize

a WSD model to extract the word sense from a lexicon. Secondly, we tackle the second problem by considering the first definition listed in Wiktionary as the basic definition. This choice aligns with the dictionary's recommendation to utilize the logical hierarchy of word senses (Babieno et al., 2022). The explicit contrast between the basic and word sense definitions corresponds better to steps 3 to 5 outlined in the flowchart of the MIP procedure (see Figure 1), as we elaborate on in the following section.

3. Methodology

In this section, we present the methodology used to develop and train ContrastWSD. Figure 2 provides an illustration of the data augmentation process and the subsequent metaphor detection process. We commence by introducing the datasets utilized in the study, followed by an overview of the word-sense augmentation procedure. Subsequently, we outline the modifications that were made to the MelBERT model's structure to enhance metaphor detection.

3.1. Data Augmentation

A major contribution of our research is the adherence to the systematic approach of the Metaphor Identification Procedure for detecting linguistic metaphors in the VUA datasets. The MIP procedure (as outlined in Figure 1) involves: (1) comprehending the general meaning of the text, (2) determining lexical units, (3) identifying the contextual meaning of the units, (4) and verifying if there is a more basic meaning. (5) If the contextual meaning deviates from the basic meaning but remains understandable by comparison, (6) the unit is labeled as metaphorical.

To align with MIP, we augmented the existing datasets used in the MelBERT model through a two-step procedure. Firstly, we employed a BERT WSD model fine-tuned on a sequence-pair ranking task (Yap et al., 2020) to extract the word sense contextual definition of the target word from WordNet. To retrieve the contextual word sense, we feed a sentence $S_c = (w_1, ..., \texttt{[TGT]}, w_t, \texttt{[TGT]}, ..., w_n)$ to the WSD model, where `[TGT]` is a special token marking the location of the target word w_t. The WSD model then performs gloss selection from WordNet and chooses the best definition D_c of w_t that fits the context. Secondly, we retrieved the basic definitions from the datasets compiled by Babieno et al. (2022). The authors selected the first definition listed in the Wiktionary dictionary as the basic definition, following the dictionary's recommendation to utilize the logical hierarchy of word senses in their guidelines.

3.2. Model Structure

Our approach to metaphor detection, involves treating it as a binary classification problem based on the target word in a given sentence $S = (w_1, ..., w_t, ..., w_n)$. Our aim is to predict whether the target word w_t is being used metaphorically or literally in the sentence. To accomplish this, we leverage the contextual and basic meanings of the target word in the given sentence. Our model utilizes three separate RoBERTa models to encode the sentence S, the isolated target word w_t, as well as the contextual and basic definitions D_c and D_b.

Following the MelBERT design, we modify the sentence S by appending the POS tag to its end, and we enclose it with two segment separation tokens `[SEP]`. Additionally, we employ three types of extra embeddings: (1) The target embedding, used to indicate the target word. (2) The local context embedding, which marks either the words in the clause containing the target word between two commas or the definition and word sense. (3) The POS embedding, used to mark the position of the POS tag. We incorporate the target word and the definitions by prepending the word to the definitions we retrieve from WordNet (for the word sense) or Wiktionary (for the basic definition). This format is similar to how words appear in a dictionary, and we do this to utilize their hidden representation later in the model. We separate the target word from the definitions using the segment separation token `[SEP]`.

The RoBERTa models produce the hidden representations $\mathbf{h}_S$, $\mathbf{h}_c$, and $\mathbf{h}_b$ encoding the sentence, the contextual definition, and the basic definition, respectively with the extensions as described. $\mathbf{h}_c$ and $\mathbf{h}_b$ are produced by averaging the embedding of all the output tokens.

$$
\begin{aligned}
(\mathbf{e}_0, ..., \mathbf{e}_t, ..., \mathbf{e}_n) = \mathbf{h}_S &= \texttt{RoBERTa}(S) \\
\mathbf{h}_c &= avg(\texttt{RoBERTa}(D_c)) \\
\mathbf{h}_b &= avg(\texttt{RoBERTa}(D_b))
\end{aligned}
$$

Since the MIP procedure focuses on the semantic contrast between the basic and contextual meanings of a word, we encode the basic and contextual meanings. Thus, our MIP layer uses the word sense embedding $\mathbf{h}_c$ and the basic definition embedding $\mathbf{h}_b$. We also use cosine similarity to measure the semantic gap between the embeddings, similar to the approach employed by Babieno et al. (2022).

$$
\mathbf{h}_{MIP} = l(\mathbf{h}_c \oplus \mathbf{h}_b \oplus CosSim(\mathbf{h}_c, \mathbf{h}_b)) \quad (1)
$$

Where $l(.)$ is a fully-connected layer. We have also introduced two additional helper layers to our model. The first layer learns the relationship between the target word's contextual embedding $\mathbf{e}_t$ and the target word embedding adjacent to the word's sense

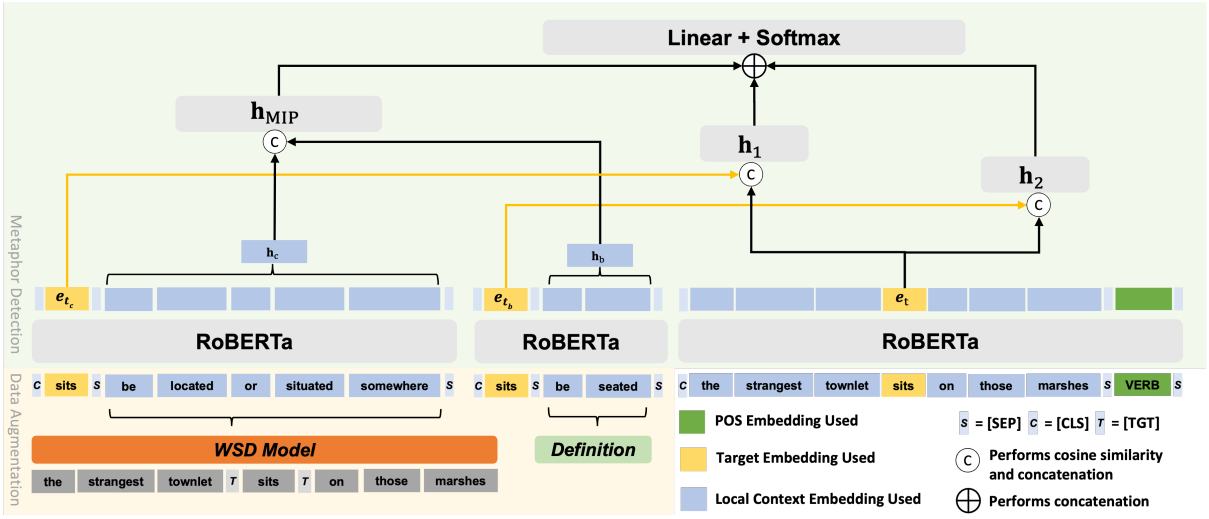

Figure 2: ContrastWSD overall framework showing both stages: (i) the data augmentation stage and (ii) the metaphor detection stage.

$\mathbf{e}_{t_c}$, while the second layer learns the relationship between the target word's contextual embedding $\mathbf{e}_t$ and the target word embedding adjacent to the word's basic definition $\mathbf{e}_{t_b}$. We believe that these helper layers will assist the MIP layer when the WSD model fails to distinguish between the word sense and the basic definition, particularly in the case of detecting novel metaphors that lack multiple definitions in the dictionary.

$$\mathbf{h}_1 = l(\mathbf{e}_{t_c} \oplus \mathbf{e}_t \oplus CosSim(\mathbf{e}_{t_c}, \mathbf{e}_t))$$
$$\mathbf{h}_2 = l(\mathbf{e}_{t_b} \oplus \mathbf{e}_t \oplus CosSim(\mathbf{e}_{t_b}, \mathbf{e}_t))$$

We concatenate the hidden vectors from the MIP and the two helper layers before feeding them to the final binary classification layer for metaphor prediction.

4. Experiments

In this section, we present the datasets, baseline models, and experimental setup used to evaluate our model. We also discuss various aspects of our experimentation process.

Datasets: Our study utilized the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010) in five pre-processed versions: **VUA-18** (Leong et al., 2018), **VUA-20** (Leong et al., 2020), and **VUAverb**, which focuses exclusively on verb metaphors. Additionally, we used **VUA-MPD-All**, which is a subset that focuses on the words that are available in WordNet, and **VUA-MPD-Conv**, a subset that focuses on conventional metaphors (Maudslay and Teufel, 2022). Each dataset comprises sentences with a labeled target word, categorized as either metaphorical or not, along with a corresponding POS tag for the target word. For each dataset, there are separate train and test datasets. The VUA-18 dataset further includes four testing subsets, each corresponding to a different POS: **verb, noun, adjective, and adverb**. We augmented the datasets with word senses extracted from WordNet using the BERT-based WSD model, and we included the basic definitions extracted from Wiktionary, as explained in the previous section.

However, we encountered instances where the WSD model failed to find a word sense or cases where there was no definition available in the dataset. To ensure a fair comparison between our model and the baselines, we created two versions of the VUA-18, VUA-20, and VUAverb training and testing datasets:

1. The original datasets were retained without removing any records. In places where no word sense was produced, we substituted it with the target word itself. For these target words, our model behaves comparable to MelBERT, enabling a comparison of the target word's embedding within and outside the context. Only in this case, our model is dependent on the contextual embeddings of the target word without external knowledge.

2. We removed the records where the WSD model failed to find a word sense and marked the pruned dataset with a minus sign (-). This was done to avoid potential noise arising from incomplete information during the model's learning process and to enable fair comparisons, ensuring that the model trains only on instances where all the necessary information is available.

For the VUA-MPD training and testing datasets,

as well as the VUA-18 POS testing subsets, we used only one version which corresponds to how we handled the datasets in version 1.

Dataset	Model	Rec	Prec	F1
VUA-18	MelBERT	77.5	79.87	78.66
	MsW_cos	77.88	**80.31**	79.07
	FrameBERT	76.78	79.33	78.03
	ContrastWSD	**78.85**	80.16	**79.50**
VUA-18 (-)	MelBERT	73.34	71.44	72.35
	MsW_cos	70.52	**74.77**	72.55
	FrameBERT	70.33	73.81	72.02
	ContrastWSD	**75.23**	72.38	**73.77**
VUA-20	MelBERT	69.51	75.58	72.40
	MsW_cos	**69.98**	75.64	72.70
	FrameBERT	69.30	75.62	72.31
	ContrastWSD	69.89	**76.60**	**73.09**
VUA-20 (-)	MelBERT	74.61	71.26	72.88
	MsW_cos	77.22	68.39	72.51
	FrameBERT	75.07	71.59	73.28
	ContrastWSD	**75.57**	**72.95**	**74.22**

Table 1: Evaluation results on VUA-18, VUA-20, and their counter pruned datasets. A bold number corresponds to the best performing model, while the underlined number the second best.

Dataset	Model	Rec	Prec	F1
verb	MelBERT	80.25	71.88	75.83
	MsW_cos	**81.85**	70.91	75.97
	FrameBERT	75.11	74.00	74.55
	ContrastWSD	78.87	**75.70**	**77.25**
noun	MelBERT	66.98	73.74	(70.19)
	MsW_cos	**68.54**	73.12	(70.76)
	FrameBERT	62.57	75.35	68.35
	ContrastWSD	66.29	**76.36**	**70.97**
adverb	MelBERT	**69.20**	71.83	70.43
	MsW_cos	64.57	71.43	67.80
	FrameBERT	65.90	74.56	69.88
	ContrastWSD	68.52	**77.45**	**72.63**
adjective	MelBERT	**67.06**	65.97	66.47
	MsW_cos	66.35	64.39	65.34
	FrameBERT	64.25	68.09	66.06
	ContrastWSD	65.73	**70.83**	**68.18**

Table 2: Performance comparison by POS tags. The results in between brackets indicate no statistically significant differences compared to ContrastWSD.

Baselines: To evaluate the performance of our model, we conducted a comparative analysis with four other baseline models that share similar concepts. The first baseline is the original model, **MelBERT** (Choi et al., 2021). The second is a modified version of MelBERT called **MsW_cos** (Babieno et al., 2022), which incorporates the basic definitions. We specifically used the cosine similarity version, as it yielded mostly the best results, as reported in their paper. The third baseline model is **FrameBERT** (Li et al., 2023), which utilizes frame embeddings to detect concept-level metaphors. We applied these models to the VUA-18, VUA-20, and VUAverb datasets, as well as their corresponding pruned versions. Additionally, we evaluated these models on the POS testing subsets after training them on the VUA-18 training dataset. The fourth baseline model is **MPD_WSD** (Maudslay and Teufel, 2022), which trains a WSD model to detect conventional metaphors. As this model employs examples from WordNet and uses distinct subsets of the VUA dataset, which we refer to as (VUA-MPD), we ran our model on the same training and testing split.

Experimental Setup The experimental setup involved using the VUA-18, VUA-20, and VUA-verb datasets, along with their corresponding pruned versions, which included training, testing, and validation subsets. For each experiment, our model and the baseline models (except the MPD_WSD) were trained on the training datasets and then evaluated on the respective testing datasets (refer to Tables 1 and 3). Additionally, the models trained on the VUA-18 training dataset were tested on the verb, noun, adverb, and adjective testing subsets (refer to Table 2), while the models trained on the VUA-20 training dataset were evaluated on the VUA-verb testing dataset which is appended by a star ($\star$) sign in Table 3. We also train and test our model on the VUA-MPD datasets (refer to Table 4).

All models were trained for three epochs using a learning rate of $3e-5$, a linear scheduler with a two-epoch warmup, and a dropout ratio of 0.2. To ensure robustness, we repeated the experiments five times using seeds 1, 2, 3, 4, and 5, following the approach in Babieno et al. (2022). The evaluation results from the test datasets were averaged over the 5 runs. We set the metaphor class weight to 3 on the original datasets, while a class weight of 4 was used on the pruned datasets, addressing the imbalance within the dataset. The experiments were conducted on a 2-GPU NVIDIA Tesla V100 with 16GB memory. Our source code is available along with all the datasets on our GitHub page[1].

We presented the findings from the MPD_WSD model as reported in the paper (Maudslay and Teufel, 2022) trained on both the VUA-MPD-All and VUA-MPD-Conv subsets. Unfortunately, we encountered memory constraints that prevented us from replicating the reported results. To maintain

[1] https://github.com/melzohbi/ContrastWSD

consistency with the MPD_WSD model, which employs the BERT-base-cased model, we also developed an alternative version of our model trained on BERT-base-cased instead of RoBERTa-base appended with a (β) sign (see Table 4).

5. Empirical Results and Case Studies

Statistical Significance The objective of this analysis is to assess the statistical significance of the performance improvements observed in the ContrastWSD model. We conduct a two-tailed t-test for each dataset, comparing our model to the baseline models, setting a significance level of p = 0.05. The results indicate that the reported differences between our model and the other baseline models are statistically significant at a confidence level of 95%, with only a few exceptions. These exceptions, which are discussed later, are marked in brackets () in Tables 2 and 3. We did not analyze the significance of performance of the MPD_WSD compared to our model as we reported the results from their paper. In the tables, the results in **bold** correspond to the best performing model, while the underlined results indicate the second best performing model.

Dataset	Model	Rec	Prec	F1
VUAverb	MelBERT	**81.08**	55.24	65.57
	MsW_cos	77.88	61.49	68.68
	FrameBERT	73.33	**71.95**	**(72.62)**
	ContrastWSD	79.18	66.97	72.54
VUAverb (-)	MelBERT	**81.40**	51.27	62.87
	MsW_cos	79.26	59.46	67.92
	FrameBERT	74.63	**70.68**	**(72.56)**
	ContrastWSD	79.28	66.66	72.42
VUAverb ($\star$)	MelBERT	72.22	76.45	74.27
	MsW_cos	**75.09**	78.00	**(76.51)**
	FrameBERT	69.96	77.60	73.55
	ContrastWSD	73.81	**78.39**	76.03

Table 3: Evaluation results on VUA-verb, VUA-verb (-), and on the VUA-verb ($\star$) datasets.

Model	F1 - All	F1 - Conv
MPD_WSD	63.10	65.90
ContrastWSD (β)	73.42	72.31
ContrastWSD	**73.84**	**73.07**

Table 4: Performance comparison on the VUA-MPD-All and the VUA-MPD-Conv subsets.

Overall Results Table 1 presents a comparison of evaluation results for the VUA-18, VUA-20, VUA-18 (-), and VUA-20 (-) datasets. The ContrastWSD model consistently outperforms all baseline models in terms of F1-score on both VUA-18 and VUA-20 datasets, even though these datasets had instances where the word sense was absent. Interestingly, the improvement in F1-score for ContrastWSD compared to other models doubled on the VUA-18 (-) and VUA-20 (-) datasets. This suggests that while the presence of word sense in the dataset may have contributed to the model's superior performance, its absence, occasionally, did not hinder its overall effectiveness.

Table 2 presents the performance comparison for the verbs, adjectives, nouns, and adverbs testing subsets. Our model shows a notable advantage in detecting metaphorical adverbs, surpassing the other models by at least 2% in F1-score. Additionally, it achieves at least a 1% improvement on verbs and adjectives. Furthermore, it significantly outperforms the recent FrameBERT model by more than 2% on the noun dataset, while achieving comparable results to MsW_cos and MelBERT with at least a 0.21% increase in F1 that was not statistically significant, while showing a notable precision gain.

Table 3 presents the comparison between our model and the baseline models on the VUA-verb and VUA-verb (-) datasets. The results demonstrate that our model significantly outperformed MelBERT and MsW_cos on the small training dataset. Moreover, when the missing word senses were removed, our model's performance remained consistently better compared to MelBERT and MsW_cos. On the other hand, while FrameBERT showed an insignificant gain of performance than our model on the small training datasets, its performance significantly declined when tested with the model trained on the larger VUA-20 training dataset. This observation suggests that our model maintains good performance even with small datasets, showing no signs of overfitting or noise introduction when trained on a larger dataset. Additionally, the performance of our model improved with considerably higher precision when trained on the larger dataset.

Table 4 compares the performance on the VUA-MPD subsets used by the MPD_WSD model. We compare the F1-score reported in their paper against our model's results, revealing a significant improvement of a minimum of 10.32% F1-score gain for our models on the subset encompassing both novel and conventional metaphors, as well as a minimum of 6.41% F1-score gain in detecting conventional metaphors.

Case Studies: As shown in the results, our ContrastWSD model exhibits relatively higher gains. To exemplify instances where ContrastWSD correctly labeled examples that were incorrectly labeled by the baselines, we conducted several case studies. Table 5 presents a few cases that demonstrate the

True Label	ContrastWSD	FrameBERT	MsW_cos	MelBERT	Sentence Context	Word Sense	Basic Definition
✓	✓	×	×	×	... and pull all nuclear **plant** out of the impending sale ...	Buildings for carrying on industrial labor.	An organism that is not an animal, especially an organism capable of photosynthesis.
×	×	×	×	×	It is a good time to **plant** hardy shrubs too.	(botany) a living organism lacking the power of locomotion.	An organism that is not an animal, especially an organism capable of photosynthesis.
✓	✓	✓	×	×	... 1,500-tonne consignment of Canadian PCBs (polychlorinated biphenyls) bound for this **plant** in Gwent.	Buildings for carrying on industrial labor.	An organism that is not an animal, especially an organism capable of photosynthesis.
✓	✓	×	×	×	Stars appear and the shadows are fallin'. You can hear my **honey** callin'	A beloved person; used as terms of endearment.	A viscous, sweet fluid produced from plant nectar by bees ...
✓	✓	✓	✓	✓	... but the tops of the mountains are still golden, as though **honey** had been poured lightly over them ...	A sweet yellow liquid produced by bees.	A viscous, sweet fluid produced from plant nectar by bees ...
×	×	✓	✓	✓	... I'll be **jumping** up and down like a what d' ya call it!	Move forward by leaps and bounds.	To propel oneself rapidly upward, downward and/or in any horizontal direction ...
✓	✓	✓	✓	✓	... there are plenty of girls who would **jump** at the chance.	Enter eagerly into.	To propel oneself rapidly upward, downward and/or in any horizontal direction ...

Table 5: Examples of predictions made by ContrastWSD and the baseline models on the VUA-20 testing dataset. ✓ marks a metaphor prediction and × marks a literal prediction.

benefits of our approach, involving contrasting word senses while considering the context of the target sentence. For these case studies, we selected the models trained on the VUA-20 dataset. For each model, we chose the best performing one among the 5 seeds. The examples shown in the table are drawn from the VUA-20 testing dataset.

For instance, we observed the word "plant" mentioned 15 times in the testing dataset: 14 times in a metaphorical sense and only once in a non-metaphorical sense. Both MsW_cos and MelBERT labeled all of these occurrences as non-metaphorical. In contrast, our model correctly identified 47% of these occurrences, while FrameBERT only identified 33% correctly. The other models have only recognized the non-metaphorical instance and none of the metaphorical instances correctly. Three of these examples are mentioned in the table.

Another example involves the word "honey", which was mentioned twice as a metaphor in the testing dataset. In the first instance, it was used in a conventional way, and our model correctly annotated it as metaphorical, leveraging the extracted word sense. The other models did not recognize this metaphorical usage. In the second instance, "honey" appeared as a novel metaphor where our model, along with the other baseline models, marked it as metaphorical. Even though the word sense was similar to the basic sense, our model still identified it as metaphorical. This indicates that our model can recognize both novel and

conventional metaphors.

Finally, the word "jump" occurs two times in the testing dataset, appearing in two different tenses and senses. In one instance, the word "jumping" was used in the literal sense, and our model correctly identified it as literal, considering that the word sense was similar to the definition. However, the other models did not recognize it as such. In the other occurrence, "jump" was used metaphorically, and our model, along with the other models, correctly identified it as metaphorical.

6. Conclusion and Future Work

In this paper, we presented a RoBERTa-based model for metaphor detection that follows the Metaphor Identification Procedure by utilizing a WSD model to extract and contrast the contextual meaning with the basic meaning of a target word. We evaluated our model on several benchmark datasets and demonstrated that leveraging senses and contrasting them can enhance the performance of metaphor detection models. Our proposed model outperformed other state-of-the-art metaphor detection models. Our work provides compelling evidence for further exploration of the use of WSD models and sense-contrasting techniques to enhance the performance of metaphor detection models.

In future work, we plan to investigate the integration of commonsense models such as COMET (Bosselut et al., 2019) to extract and utilize the common sense knowledge from the target word. COMET was used extensively in recent metaphor generation models (Elzohbi and Zhao, 2023). We believe that this integration will enable better differentiation of novel metaphors from nonsensical expressions.

7. Bibliographical 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.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics (ICCL)*, pages 86–90.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Luana Bulat, SC Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. Association for Computational Linguistics.

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243.

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*, pages 1763–1773.

Mohamad Elzohbi and Richard Zhao. 2023. Creative data generation: A review focusing on text and poetry. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC)*, pages 29–38.

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

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.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. Sensebert: Driving some sense into bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loïc Barrault. 2023. Framebert: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1550–1555.

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)*, pages 1222–1231.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics (ICCL)*, pages 65–77.

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.

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

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)*, pages 4240–4251.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, et al. 2010. A method for linguistic metaphor identification. *Converging Evidence in Language and Communication Research. John Benjamins, Amsterdam*, 14.

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 (ACL)*, pages 30–39.

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*, pages 1971–1981.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with cnn-lstm model. In *Proceedings of the workshop on figurative language processing*, pages 110–114.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020 (ACL)*, pages 41–46.