How Can Metaphor not Handle Anomaly? Metaphor Detection with Anomalous Text

Anonymous ACL submission

Abstract

Metaphor is essentially literal shifts in meaning, which is manifested as a mismatch between the literal meaning of the target word and its contextual context. In metaphor research, the theory of *Selection Preference Violation (SPV)* is commonly used to identify metaphor, which the target word occurs less frequently in the surrounding words in its context, yielding a mismatch. Researchers are mainly concerned with considering such collocational mismatch as a metaphorical expression, yet they tend to overlook that collocational mismatch may also be a syntactic anomaly. We integrate syntactic anomaly into the metaphor detection. First, we use ChatGPT to construct a dataset containing syntactic anomaly, called the LMA. Second, we propose a model for enhanced literal, metaphor, and syntactic anomaly detection (MetaLA), which considers not only the target word and context in classification detection, but also adds other semantic contexts to reduce misclassifying anomaly as metaphor. We explore the relationship between literal, metaphor and syntactic anomaly, as well as the role of introducing SPV. Our experimental results show that syntactic anomaly reduce the model's correctness for metaphor detection, and that SPV reduces this correctness even further. Finally, we compare MetaLA with existing metaphor detection methods as well as other large language models (LLM) to demonstrate the effectiveness of our approach in literal, metaphor and syntactic anomaly detection.

1 Introduction

Metaphor is a rhetorical expression that, from a linguistic point of view, is a universal linguistic expression that represents other concepts (Lagerwerf and Meijers, 2008). In a given context, metaphor is the use of one or more words to imply another concept, rather than adopting its literal meaning directly (Fass, 1991). For example, in the sentence "This task is like a bottomless pit!" the metaphorical meaning of the word "bottomless pit" is "a difficulty or challenge that has no end or solution" rather than a literal "underground cavern". This suggests that metaphor detection requires an understanding of the metaphorical expression and its relationship to the contextual word. Since metaphor play a key role in cognitive and communicative functions, this is likely to benefit many Natural Language Processing (NLP) tasks such as sentiment analysis (Cambria et al., 2017; Li et al., 2023a), communication platform (Dybala and Sayama, 2012), and psychological security (Riloff et al., 2018).

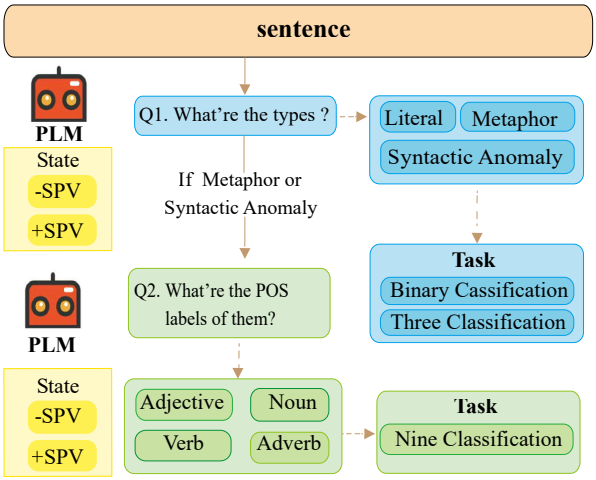


Figure 1: Mission description. Pre-trained Language Model (PLM) is required to recognize and classify literal, metaphor, and syntactic anomaly. The SPV is used to detect whether there is a relationship violation between the target word and the context word of a sentence.

In metaphor detection tasks, previous studies generally choose to use *Selection Preference Violation (SPV)* recognition methods (Wilks, 1975, 1978; Mao et al., 2019). They recognizes metaphor by identifying the relationship between the target word and the context word. Let us consider an ex-

ample: "My computer chews on wires", the word "chews" is considered metaphorical. Because in the context of "computer" and "wires", the act of "chews" is unusual. The "computer" can't chew, and "wires" don't have the ability to chew food. Consider another example "The girl comforted the clock.". The "girl" is alive, while "clock" is inanimate. Therefore, "clock" is not an appropriate argument for "comforted". This example belong to a mismatch of verb and object and is non-metaphorical. While previous researchs focus mainly on considering collocational mismatch as a metaphorical expression, they often overlook the fact that collocational mismatch can also be a manifestation of syntactic anomaly.

In NLP, syntactic anomaly is among the common types of anomaly (Lunsford and Lunsford, 2008). Syntactic anomaly is mainly found in grammatical, which is manifested as irregularities in sentence structure, non-compliance with grammatical rules, or deviations from usual linguistic expressions. Metaphor is essentially literal deviations with collocational anomaly, i.e., unusual combinations between literal meanings and meanings of other words. Metaphor detection systems often incorrectly recognize syntactic anomaly as metaphor. However, no one has yet specifically linked metaphor to syntactic anomaly.

In this paper, syntactic anomaly detection is introduced into the metaphor detection task from the perspective of dealing with syntactic anomaly and metaphor (see figure 1). We applied Chat-GPT to modify some of the literal sentences of VUA to construct a syntactic anomaly dataset (LMA). In addition, we propose MetaLA, which is an approach that utilizes multiple theories of metaphor detection. The MetaLA's components mainly include the dual encoder, SPV, and the Metaphor Detection Enhancement Module (MDEM). The principle of MDEM is to utilize MIP and MIPVU for interactive computation. Where MIP identifies metaphor by analyzing the contextual and literal meanings of target words, MIPVU extends and improves on MIP by introducing more conceptual metaphor analysis. Therefore, we consider MIP as a basic metaphor recognition framework, and then utilize MIPVU for further analysis and understanding. By introducing both MIP and MIPVU into metaphor detection, we can guide the model to focus more on distinguishing metaphor and reduce the misclassification of syntactic anomaly sentences.

In summary, our contributions are as follows:

1. We are the first work to focus on the relationship between metaphor and syntactic anomaly, introducing syntactic anomaly detection as part of the metaphor detection task.

2. We have successfully constructed a dataset containing syntactic anomaly (LMA). This dataset can be combined with datasets from other tasks and used to improve the performance of the model in handling syntactic anomaly.

3. We propose MetaLA, a model that effectively distinguishes between literal, metaphor and syntactic anomaly sentences.

4. We provide the first insight into the role of SPV for metaphor detection and syntactic anomaly detection. Our ablation experiments show that the performance of the model degrades when using SPV for the detection of metaphorical and syntactic anomaly.

2 Related Work

2.1 Metaphor Detection

Metaphor detection is a sequence annotation task that aims to determine whether a target word is a metaphorical expression in context, with "1" being metaphorical and "0" being non-metaphorical. Current metaphor detection tasks focus on supervised methods. For example, Mao et al. (2019) directs the model to compare the underlying and contextual meanings of target words to determine metaphor, and Le et al. (2020) uses a textual dependency tree structure to construct metaphor. Li et al. (2023b) uses two encoders, one of which is fine-tuned by FrameNet (Fillmore et al., 2002). Choi et al. (2021) is similar to (Mao et al., 2019) but replaces the LSTM model with RoBERTa. Zhang and Liu (2022) introduced example sentences as a control, using literal meaning samples from the original dataset. Su et al. (2020) used a cueing approach to translate metaphor detection into reading comprehension and introduced local textual information. Su et al. (2021) based on MIP and SPV modeling, respectively, with a simple DNN architecture for MIP and a novel multi-head contextual attention mechanism for SPV, designed to perform metaphor recognition for end-to-end sequences. Furthermore, Pramanick and Mitra (2018)

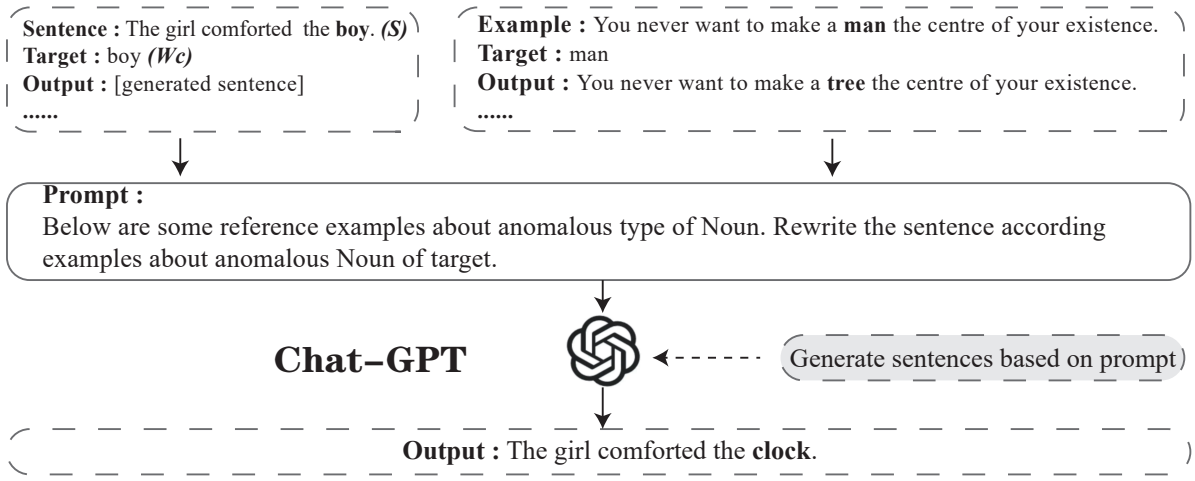


Figure 2: Syntactic Anomaly data generation. We take the example of generating noun syntactic anomaly data. The input consists of a correct literal sentence ***(S)*** and the target word of the sentence ***(Wc)***. We have designed a prompt containing specific examples. Guided by this prompt, ChatGPT needs to perform the same lexical modifications on the input sentence to generate the corresponding noun syntactic anomaly data.

proposed an unsupervised framework for recognizing metaphorical adjective-noun word pairs using cosine similarity as well as derivatives of abstraction ratings and edit distances for clustering, and evaluated it on a large TSV dataset. Badathala et al. (2023) introduces exaggerated corpus knowledge into metaphor detection, while Zhang and Liu (2023) uses adversarial learning to guide the model in learning data distributions across multiple tasks.

2.2 Anomaly Detection

Anomaly detection is an important aspect of text processing. In the field of NLP, syntactic anomaly account for a relatively large number of anomaly problems, including lexical mismatch (Lunsford and Lunsford, 2008). Lunsford and Lunsford (2008) continues to study text anomaly types based on the previous work and summarizes a list of anomaly. Common types of textual anomaly are wrong sentence structure, such as lack of subject and verb agreement. Søby et al. (2023) focuses on the types of syntactic anomaly as well as the frequency of anomaly in Danish written expressions, etc., involving various subtypes (word order errors, verb consistency errors). Bock and Miller (1991) point out that speakers may commit subject-predicate agreement errors when singular nouns are followed by plurals. Nicol et al. (1997) further investigate this anomaly in (Bock and Miller, 1991). Barton and Sanford (1993); Nieuwland and Van Berkum (2006), study the problem of local incoherence (verb-object violation) in texts such as "Tom drinks the sunshine every morning". Nieuwland and Van Berkum (2006) favors the study of syntactic anomaly with and without vital violations. Ni et al. (1998) explore how the parser responds to explicit sentences containing both syntactic and pragmatic anomaly. In addition, Herbelot and Kochmar (2016) focus on the adjective-noun combination anomaly (... My friends have a hard time calling me on a classical phone ...). Similarly, Vecchi et al. (2011) applied some combinatorial models to detect adjective-noun combinations with semantic syntactic anomaly.

3 Method

3.1 Mission Description

In this paper, syntactic anomaly detection is introduced based on the metaphor detection task. Regarding metaphor detection, most of the previous studies use sentence-level labeling methods (Mao et al., 2019; Le et al., 2020; Su et al., 2020; Choi et al., 2021). And syntactic anomaly (Ivanova et al., 2012, 2017) is generally studied at the sentence level as well. We emphasize the classification of metaphor detection and syntactic anomaly detection at the sentence level (See Figure 1).

3.2 MetaLA

In this paper, we propose an approach, MetaLA, for literal, metaphor and syntactic anomaly detection (see figure 3). The architecture of MetaLA consists of an encoder for sentences and target word, SPV, and a Metaphor Detection Enhancement Module

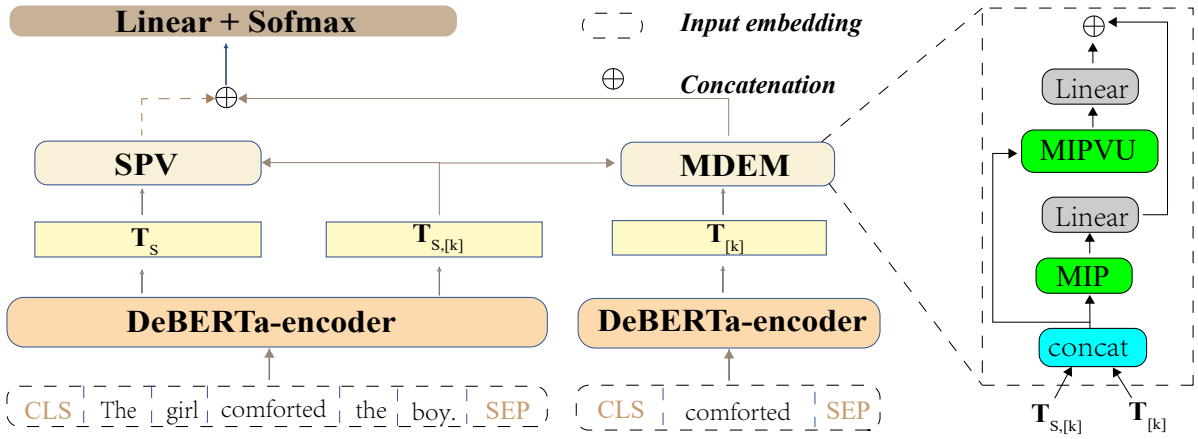


Figure 3: MetaLA model architecture. The DeBERTa-encoder embeds the input sentences with the target word and output the average pooled result. The "- - →" indicate that SPV is introduced in the experiment.

Below are some reference examples about anomalous type of adjective. Rewrite the sentence according examples about anomalous adjective of target.
Example 1: It was very difficult for my friends to call me with the small phone. **Target**: small **Output**: It was very difficult for my friends to call me with the delicious phone...
Example 2: We need to consider the availability of large databases for hazard studies. **Target**: large **Output**: We need to consider the availability of delicious databases for hazard studies.
Example 3: Manuals which may contain maps, schematic diagrams, and other materials warrant separate consideration. **Target**: schematic **Output**: Manuals which may contain maps, nervous diagrams, and other materials warrant separate consideration.
Example 4: Early in the morning, the sunlight pours in the quiet garden. **Target**: quiet **Output**: Early in the morning, the sunlight pours in the delicate garden.
Sentence: Oh dear, Miss Williams said on an indrawn breath. **Target**: indrawn **Output**: [generated sentence]

Table 1: Demonstration of prompt generated for syntactic anomaly for the ChatGPT. We take adjective anomaly as an example for the demonstration and set up four sets of examples.

(MDEM). Given sentence input and target word input, DeBERTa-encoder converts each word in the sentence as well as the target word into each token. In addition, the position of each word and target word will be embedded. We also use a special categorization token "CLS" for sentence starters and a special separation token "SEP".

In MDEM, we take the *Metaphor Identification Theory (MIP)* and its extension *(MIPVU)* as the core part. MIP indirectly identifies metaphor and can only determine whether a word is a metaphor or not by analyzing the contextual and literal meanings of the target word. MIPVU, on the other hand, extends and improves on MIP by introducing more conceptual metaphor to be analyzed. Compared to MIP, MIPVU not only extends non-metaphorically related words, but also considers multiple types of metaphor (e.g., direct, indirect, and the metaphorical flag MFlag) (Steen et al., 2010). For direct metaphors (e.g., like and as) and indirect metaphor, they can then be analyzed and identified using conceptual metaphor. Therefore, we consider MIP as a basic metaphor identification framework, which is then further analyzed and understood using MIPVU.

Given a sentence $S = \{v_0, ..., v_k, ..., v_{n+1}\}$, where v_0 = CLS, v_{n+1} = SEP. We use DeBERTa-encoder to encode the input sentence S and convert it into an embedding vector $\{T_0, ..., T_{s,[k]}, ..., T_{n+1}\}$ for computation.

$$T_s = \text{DeBERTa-Enc}(\{v_0, ..., v_k, ..., v_{n+1}\}) \quad (1)$$

Here, we use T_0 to denote the embedding vector of CLS, and T_{n+1} to denote the embedding vector of SEP. While $T_{s,[1]}, ..., T_{s,[n]}$ are the hidden

state outputs of each marker, we consider them as contextual embeddings, including the contextual embedding of the target word $T_{s,[k]}$ $(1 \leq k \leq n)$. Moreover, T_k denotes the literal meaning of the target word.

$$T_k = \text{DeBERTa-Enc}(\{v_k\}) \quad (2)$$

After obtaining the sentence embedding T_s, the target word context embedding $T_{s,[k]}$, and the literal embedding T_k, we use them as inputs to SPV, MDEM, and model them to obtain the hidden vectors H_S and H_M.

$$H_S = f_1([T_s; T_{s,[k]}]) \quad (3)$$

$$H_M = f_{mip}([T_s; T_{s,[k]}]) * f_{mipvu}([T_s; T_{s,[k]}]) \quad (4)$$

Finally, we utilize H_S and H_M to compute the prediction scores for classification.

4 Prompt Construction

We design a prompt whose process consists of being given a set of prompt examples. Each example contains the original sentence, the target word, and the generated output sentence. The prompt consist of a task description title and an example composition. To ensure the rigor of syntactic anomaly data generation, we refer to the current literature based on GPT-3 prompt (Yoo et al., 2021; Reynolds and McDonell, 2021; Chakrabarty et al., 2023) when developing our prompt. For each syntactic anomaly type, we customize different task descriptions and examples to make them more relevant. Table 1 shows our specific prompt (in the case of adjective). Other types of syntactic anomaly sentences are generated in a similar way.

5 Dataset

This section delves into the construction of the syntactic anomaly dataset (LMA).

5.1 Metaphor Datasets

5.1.1 VUAMC

The VUAmsterdam Metaphor Corpus [1](Steen et al., 2010) metaphorically annotates each lexical unit (187,570 in total) in a subset of the British National Corpus (BNC Consortium, 2007) (Edition et al.). The corpus contains 115 texts of four different types, covering academic, conversational, fictional, and journalistic texts.

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

5.1.2 VUA ALL POS

The VUA ALL POS dataset is a key component of the metaphor detection shared task (Leong et al., 2018, 2020). In VUA ALL POS, all real words (including adjective, verb without have, do, be, noun) in a sentence are labeled. To distinguish it from the VUA ALL POS defined in (Leong et al., 2018, 2020), we name the VUA ALL POS dataset that contains both real-sense words and dummy words as VUA ALL.

5.2 Dataset Construction

We considered the key issues of quantity and distribution in depth. First, we carefully selected 3243 token-level literal sentences from VUA based on four part-of-speech (Adjective, Noun, Adverb, Verb). Second, we expanded this data and further filtered it to obtain 22,010 sample data of sentence-level literal meanings. Based on the prompt presented in the methodology of Section 4, we extracted 20% of the selected 22,010 data samples as input to ChatGPT with syntactic anomaly sentences (see Figure 2 for details). After generating all the syntactic anomaly data, we merged and adjusted them with the literal meaning data samples to form the syntactic anomaly dataset LMA. Our modified syntactic anomaly data are generally similar to natural discourse, and some of them are demonstrated in the Appendix.

6 Experiments

6.1 Baseline

BERT: BERT (Devlin et al., 2018) employs a bidirectional Transformer encoder, available in both base and large versions. In the pre-training phase, BERT performs two tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).
RoBERTa: Unlike BERT, RoBERTa removes the NSP task in pre-training, i.e., it no longer determines whether two sentences are adjacent.
ALBERT: ALBERT was proposed in 2019 by Lan et al. (2019) to improve the scalability and efficiency of BERT. The design of ALBERT consists of several versions.
DeBERTa: He et al. (2020) Improving upon BERT, DeBERTa was introduced in 2020. DeBERTa incorporates a decoding-enhanced mechanism and disentangled attention, contributing to enhanced performance.
MeBERT_MIP: MeBERT_MIP is one of the metaphor detection models in (Choi et al., 2021).

Model	M-L			SA-L			T-C			N-C		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT-bs	0.756	0.740	0.744	0.749	0.750	0.749	0.746	0.730	0.734	0.728	0.751	0.729
BERT-lg	0.750	0.762	0.754	0.747	0.758	0.751	0.741	0.739	0.740	0.738	0.730	0.733
ALBERT-bs	0.744	0.755	0.746	0.736	0.742	0.739	0.726	0.731	0.725	0.734	0.697	0.710
ALBERT-xxlg	0.778	0.783	0.780	0.769	0.759	0.761	0.767	0.751	0.752	0.751	0.730	0.739
RoBERTa-bs	0.770	0.772	0.769	0.762	0.770	0.764	0.758	0.768	0.761	0.751	0.760	0.754
RoBERTa-lg	0.773	0.779	0.772	0.767	0.778	0.770	0.769	0.756	0.763	0.765	**0.777**	0.755
DeBERTa-bs	0.758	0.760	0.759	0.754	0.759	0.755	0.748	0.745	0.745	0.745	0.734	0.739
MelBERT_MIP	0.776	**0.795**	0.779	0.766	0.759	0.765	**0.773**	0.747	0.759	0.718	0.736	0.725
MetaLA (our)	**0.783**	0.790	**0.785**	**0.781**	**0.788**	**0.781**	0.762	**0.778**	**0.770**	**0.766**	0.760	**0.764**

Table 2: Model performance without the introduction of SPV. The "**bs**" stands for the "base" version of the baseline model, and the "**lg**" denotes the "large" version of the baseline model. Our experiments include metaphor-literal detection (**M-L**), syntactic anomaly-literal detection (**SA-L**), three-classification (**T-C**) and nine-classification detection (**N-C**). The evaluation metrics include precision (**Prec**), recall (**Rec**) and composite metric (**F1**), where F1 is the core metric (best in **bold**).

This model mainly utilizes context words as well as the Metaphor Identification Process (MIP) to identify whether the target word is metaphor or not.

6.2 Experimental Design

We combine the syntactic anomaly dataset LMA and the metaphor dataset VUA to conduct two sets of experiments, each covering the same fine-grained sub-experiments.

Experiment 1: The sub-experiments of Experiment 1 covered binary classification, three classification, and nine classification detection. Specifically, binary classification detection includes metaphor-literal detection, syntactic anomaly-literal detection, and metaphor-syntactic anomaly detection. Three classification detection covers literal-metaphor-syntactic anomaly detection. The nine classification detection, on the other hand, provides a more detailed division of metaphor and syntactic anomaly based on pos labels (Adjective, Noun, Verb, Adverb) on the basis of the three classification detection.

Experiment 2: In the second set of experiments, we introduced SPV, combining it with the baseline model to form a new baseline model. The purpose of Experiment 2 is to investigate whether SPV has an impact on the model's performance in literal, metaphor, and syntactic anomaly classification tasks.

7 Implementation

In both sets of experiments, our experimental setup is similar to (Choi et al., 2021). The learning rate is initialized to 3e-5,warmupepoch is set to 3. The learning rate is controlled by a linear warmup scheduler, and the learning rate is gradually increased during the warmup period. In addition, we set the dropout rate to 0.2. The hidden layer of the classifier is set according to the size of the model, which is set to 768 for the base model and 1024 for the large model. The maximum number of training rounds is set to 20. The K-fold cross-validation is set to 10. The maximum length of the sentence is limited to 150 Both experiments were run on a cloud server equipped with a single A100 80G GPU.

8 Experimental Results

8.1 Results and Analysis

Here we will compare and analyze the results of Experiments 1 and 2 both horizontally and vertically.

In Experiment 1, we evaluated the performance of the baseline model and MetaLA, and the specific results are shown in Tables 2 and 4. Observing the metric scores of metaphor-literal detection and syntactic anomaly-literal detection in these two tables, we can find that these models not only perform well in metaphor recognition, but also achieve better performance in syntactic anomaly recognition. Compared to the other pre-trained models,

Model	M-L			SA-L			T-C			N-C		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT-bs*	0.744	0.750	0.748	0.762	0.746	0.756	0.727	0.729	0.727	0.718	0.715	0.713
BERT-lg*	0.733	0.766	0.758	0.755	0.737	0.753	0.745	0.726	0.733	0.720	0.735	0.726
ALBERT-bs*	0.752	0.761	0.755	0.754	0.736	0.744	0.731	0.713	0.719	0.719	0.697	0.705
ALBERT-xxlg*	**0.797**	0.773	0.789	0.764	0.761	0.763	0.752	0.742	0.746	0.711	0.730	0.723
RoBERTa-bs*	0.784	0.753	0.776	0.757	0.773	0.767	0.727	0.740	0.732	0.710	0.724	0.718
RoBERTa-lg*	0.782	0.775	0.775	0.763	0.781	0.773	0.754	0.767	0.756	0.723	0.744	0.730
DeBERTa-bs*	0.756	0.769	0.765	0.762	0.767	0.763	0.737	0.748	0.740	0.750	0.731	0.738
MelBERT_MIP*	0.787	0.766	0.782	0.771	0.762	0.771	0.730	0.739	0.734	0.742	0.724	0.728
MetaLA* (our)	0.785	**0.796**	**0.794**	**0.785**	**0.793**	**0.785**	**0.777**	**0.785**	**0.759**	**0.770**	**0.736**	**0.744**

Table 3: Model performance when introduce the SPV. The experiments include metaphor-literal detection (**M-L**), syntactic anomaly-literal detection (**SA-L**), and three-classification detection (**T-C**) (best in **bold**).

Model	M-SA		
	Prec	**Rec**	**F1**
BERT-bs	0.746	0.740	0.741
BERT-lg	0.738	0.745	0.746
ALBERT-bs	0.735	0.734	0.734
ALBERT-xxlg	0.756	0.764	0.758
RoBERTa-bs	0.762	0.765	0.762
RoBERTa-lg	0.755	0.772	0.768
DeBERTa-bs	0.753	0.747	0.749
MelBERT_MIP	0.748	0.766	0.763
MetaLA (our)	**0.776**	**0.784**	**0.778**
BERT-bs*	0.739	0.752	0.734
BERT-lg*	0.733	0.746	0.738
ALBERT-bs*	0.728	0.717	0.722
ALBERT-xxlg*	0.725	0.735	0.729
RoBERTa-bs*	0.738	0.750	0.742
RoBERTa-lg*	0.733	0.748	0.747
DeBERTa-bs*	**0.759**	0.738	0.743
MelBERT_MIP*	0.741	0.751	0.744
MetaLA* (our)	0.755	**0.768**	**0.762**

Table 4: Model performance on metaphor-syntactic anomaly (**M-SA**) with and without SPV.

our MetaLA achieves the highest scores on the F1-scores, which are 0.785, 0.781, and 0.778, respectively. This indicates that the dataset LMA is effective. In the three-classification sub-experiment, the performance of the eight baseline models generally decreased slightly, with MetaLA achieving the highest F1-score of 0.783. Looking deeper into the results of the nine-classification experiments, we found that the performance of the baseline models decreased further compared to the three-classification experiments. This may be due to the fact that in the nine-classification sub-experiment, the model needs to further differentiate pos types for metaphors and syntactic anomalies, leading to an increase in classification difficulty. However, MetaLA still outperforms other methods and pre-trained models.

In Experiment 2, we explored the effect of SPV on metaphor and syntactic anomaly recognition, and the specific experimental results are presented in Tables 3 and 4. After comparing with the F1 scores in Experiment 1, we found that the performance metrics of the baseline model improved in the metaphor-literal and syntactic anomaly-literal tasks. In particular, our model metaLA achieved the highest level with scores of 0.794 and 0.785. However, the performance of the model declined in the metaphor-syntactic anomaly sub-experiment, where the F1 score for RoBERTa-large dropped to 0.747, a decrease of 0.021. Notably, in the three and nine classification experiments, we observed that the performance of the model did not improve, but rather showed a decreasing trend. The results of Experiment 2 suggest that the introduction of SPV in multitasking scenarios may lead to complex cross-influences that increase the model's confounding of metaphorical and syntactic anomalies. In this case, metaLA induces the model to focus more on metaphorical features, further improving the performance of metaphor recognition.

Accurate identification of metaphor and syntactic anomaly is crucial. The syntactic anomaly do not only exist in metaphor detection, but also cover other NLP tasks. Our research can provide assis-

State	Sentence	M	S-A	L
-SPV	What are you going to cook, Miss Mair?			✓✓
	we need to consider the availability of delicious databases.		✓✓	
	It's my life that is about to go down the plughole.	✓✓		
	It was very difficult for my friends to call me with the delicious phone...		✓✓	
	She gave him a beautiful smile, which lit her lovely face.	✓	✓	
	Right we go across the cat now nicely.	✓	✓	
	You could feel the atmosphere when you were sat in the car.	✓✓		
+SPV	What are you going to cook, Miss Mair?			✓✓
	we need to consider the availability of delicious databases.		✓✓	
	It's my life that is about to go down the plughole.	✓✓		
	It was very difficult for my friends to call me with the delicious phone...	✓	✓	
	She gave him a beautiful smile, which lit her lovely face.	✓	✓	
	Right we go across the cat now nicely.	✓	✓	
	You could feel the atmosphere when you were sat in the car.	✓	✓	

Table 5: Case study examples. We provide several examples of sentences with literal, metaphor, and syntactic anomaly that are distinguished using the BERT-base model. The last three columns of the table show the range of recognition results from the model. The "M" stands for metaphor, "S-A" for syntactic anomaly, and "L" represents literal. The symbol "✓" express the correct labeling result of the corresponding sentence, while the "✓" indicate the recognition result of the model.

tance to the field of NLP by more accurately identifying and distinguishing syntactic anomaly.

8.2 Results Visualization

We visualize the results of experiment to visually compare the performance of the models. We focused on the core metric F1 score. First, we performed a side-by-side comparison visualization showing the performance of all models on different classification tasks (as shown in Figure 4 (a)). We can find that the folds gradually decrease in the direction from binary-classification, three-classification to nine-classification tasks. Second, we performed a longitudinal comparison visualization of individual models on different tasks (as shown in Figure 4 (b)). Looking at each subplot, we find that the model introducing SPV outperforms the case without introducing SPV on the M-L and SA-L tasks, while the opposite is true on the other tasks. The visualization results show a gradual decrease in model performance as task complexity as well as SPV increases.

8.3 Case Study

We provide several literal, metaphor, and syntactic anomaly sentences and use them as model inputs to the BERT-base model and the three classification detection task, as shown in Table 5. From the table, we observe that without introducing SPV, the model incorrectly recognizes some metaphor and syntactic anomaly sentences. Further observing the model performance after the introduction of SPV, we find that the model misidentification is further exacerbated by misidentifying more syntactic anomaly sentences as metaphor. This suggests that SPV exacerbates the model's confusion between metaphors and syntactic anomaly.

9 Conclusion

In metaphor detection tasks, collocations are not only metaphor, but may also be syntactic anomaly. We constructed a high-quality syntactic anomaly dataset LMA. Based on LMA, we also explore the role of SPV for metaphor and syntactic anomaly detection. In addition, for literal, metaphor and syntactic anomaly classification, we propose a method to detect them, i.e., MetaLA. MetaLA guides the model to be able to focus more on metaphorical features to classify them, and achieves good results. To the best of our knowledge, this paper is the first one devoted to syntactic anomaly and metaphor tasks. The experimental results show that there is a large confusion between metaphor and syntactic anomaly in the model, which is exacerbated by the introduction of SPV.

10 Limitations

In this study, we propose a task that specializes in anaphora and syntactic anomaly. We use ChatGPT in constructing the anomaly data. Despite the high performance of ChatGPT, there are some discrepancies. It is possible that every piece of data generated ChatGPT has some differences for our prompt, which can lead to some mislabeling of the data. And our range of anomaly data types is limited to only four types. In addition to that, we did not investigate at a finer granularity level, such as token level. In future work, we will further explore more types of syntactic anomaly and how to efficiently differentiate between metaphor and syntactic anomaly.

11 Ethics Statement

In this study, we strictly adhered to the guidelines of academic and research ethics. We place special emphasis on transparency and openness of information, and explicitly cite the public data sources cite in order to fully respect the original authors and data providers of relevant research in the field of metaphor recognition. Throughout this research, we have never intentionally and maliciously criticized or plagiarized the work of others. Our approach is fully consistent with the principles of academic integrity and aims to ensure full recognition of the work and contributions of those who have gone before us. At every step of the research process, we have kept in mind the requirements of academic ethics and are committed to ensuring the authenticity, transparency and fairness of our research. We are confident that such an attitude towards research will make a positive and sustainable contribution to the prosperity and growth of the academic community.

References

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*.

Stephen B Barton and Anthony J Sanford. 1993. A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & cognition*, 21(4):477–487.

Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive psychology*, 23(1):45–93.

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

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

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*.

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*.

Pawel Dybala and Kohichi Sayama. 2012. Humor, emotions and communication: Human-like issues of human-computer interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

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

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*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Aurélie Herbelot and Ekaterina Kochmar. 2016. 'calling on the classical phone': a distributional model of adjective-noun errors in learners' english. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 976–986.

Iva Ivanova, Holly P Branigan, Janet F McLean, Albert Costa, and Martin J Pickering. 2017. Do you what i say? people reconstruct the syntax of anomalous utterances. *Language, Cognition and Neuroscience*, 32(2):175–189.

Iva Ivanova, Martin J Pickering, Holly P Branigan, Janet F McLean, and Albert Costa. 2012. The comprehension of anomalous sentences: Evidence from structural priming. *Cognition*, 122(2):193–209.

Luuk Lagerwerf and Anoe Meijers. 2008. Openness in metaphorical and straightforward advertisements: Appreciation effects. *Journal of Advertising*, 37(2):19–30.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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, Frank Guerin, and Chenghua Lin. 2023a. The secret of metaphor on expressing stronger emotion. *arXiv preprint arXiv:2301.13042*.

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*.

Andrea A Lunsford and Karen J Lunsford. 2008. " mistakes are a fact of life": A national comparative study. *College Composition and Communication*, pages 781–806.

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.

Weijia Ni, Janet Dean Fodor, Stephen Crain, and Donald Shankweiler. 1998. Anomaly detection: Eye movement patterns. *Journal of Psycholinguistic Research*, 27:515–539.

Janet L Nicol, Kenneth I Forster, and Csaba Veres. 1997. Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, 36(4):569–587.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, 18(7):1098–1111.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. 2018. Proceedings of the 2018 conference on empirical methods in natural language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Katrine Falcon Søby, Byurakn Ishkhanyan, and Line Burholt Kristensen. 2023. Not all grammar errors are equally noticed: error detection of naturally occurring errors and implications for eye-tracking models of everyday texts. *Frontiers in Psychology*, 14.

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.

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.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9.

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.

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

Shenglong Zhang and Ying Liu. 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*.

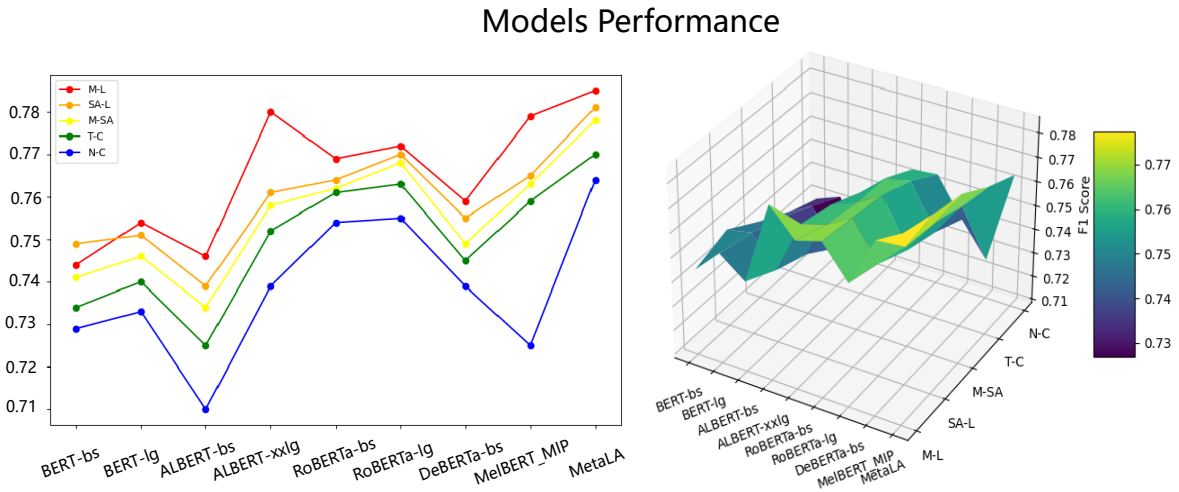
A Appendix

A.1 LMA Display

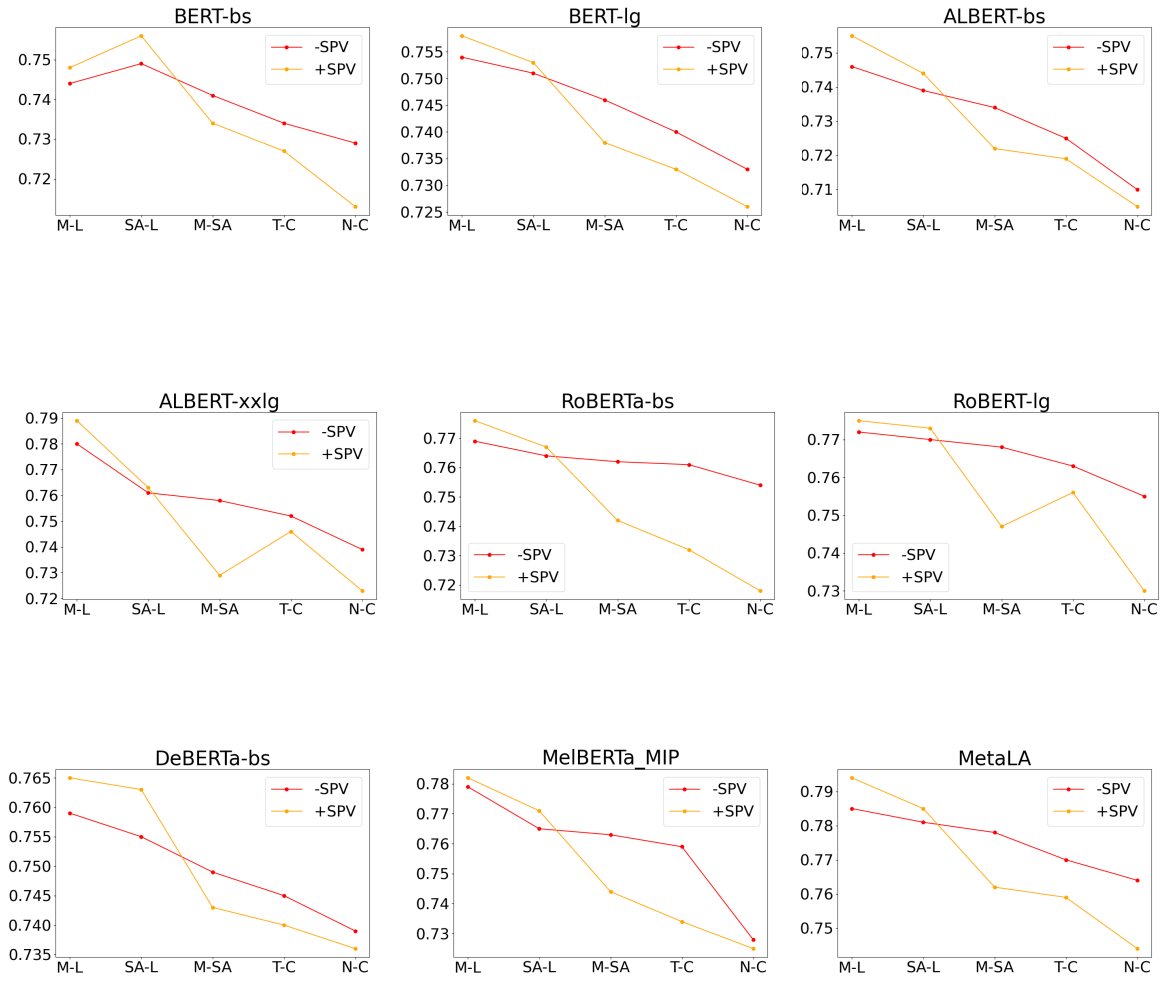
Display 1: It is a delicious challenge to Israel's occupation, conceived in the image of earlier challenges of the Intifada. **Target**: delicious **Type**: Adjective
Display 2: For the most part this is a humble assumption to make. **Target**: humble **Type**: Adjective
Display 3: Right we go across the cat now nicely. **Target**: cat **Type**: Noun
Display 4: The Government White Paper Crime, Justice and Protecting the Public propose a radical change for apple practice. **Target**: apple **Type**: Noun
Display 5:Parents wants to be part of the decision-making and did not feel that they were. **Target**: wants **Type**: Verb
Display 6: But if the weather is quite pleasant, it would be nice to open someone over, I suppose. **Target**: open **Type**: Verb
Display 7: Look at me yesterday you're talking to me I can hear you. **Target**: yesterday **Type**: Adverb
Display 8: Oh I hate those things, personally, but tomorrow, there we are. **Target**: tomorrow **Type**: Adverb

Table 6: Four syntactic anomaly types are shown (including Adjective, Noun, Verb, Adverb), each including two displays. In each display, we give the sentence, the target word, and the anomaly type.

We show some samples of our syntactic anomaly dataset, as shown in Table 6. Given that our syntactic anomaly data is modified based on the literal meaning of the VUA dataset, the generated syntactic anomaly data is modified while maintaining the original natural semantic style. Meanwhile, combined with the relevant experimental results of syntactic anomaly detection, we observe that the model performs better in performance detection on LMA, which indicates that our syntactic anomaly data modified using LLM is of high quality. We have fully considered the main types of syntactic anomaly and the number of samples of each type to ensure the uniformity of data distribution. We are confident that our dataset can contribute to the field of natural language processing.



(a)



(b)

Figure 4: (a) shows a visual comparison of the performance of different models under five classification tasks. The horizontal coordinates represent all models and the vertical coordinates indicate the corresponding F1 score. (b) depicts a comparison of the performance visualization of a single model on different tasks. The horizontal coordinates represent the categorization tasks, while the vertical coordinates represent the corresponding F1 score.