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 in 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. Syntactic anomaly are mainly found in grammatical structures or grammatical rules, which are manifested as irregularities in sentence structure, non-compliance with grammatical rules, or deviations from usual linguistic expressions. In this paper, we integrate syntactic anomaly into the study of metaphor detection. Specifically, we craft a prompt. Based on this prompt, we use GPT-3 to generate a dataset containing literal, metaphor, and syntactic anomaly, called the LMA. We test our dataset in a series of related experiments. We explore the relationship between literal, metaphor and syntactic anomal, as well as the role of introducing SPV. We provide experimental analysis.

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 utilizes one or more words to represent another concept rather than adopting the literal meaning of the expression (Fass, 1991). For example, in the "This program is a headache!", the contextual meaning of headache is "something or someone that causes worry or trouble", which is different from its literal meaning of "constant pain in the head". 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 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). Metaphor detection is challenging because it requires the model to have a deep understanding of the non-literal meanings in the text and to detect them accurately across a wide range of text types to better capture the intent of the text .

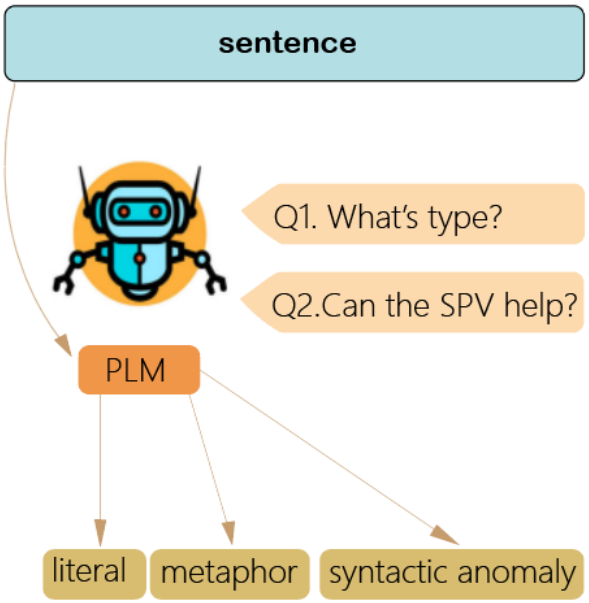

Figure 1: Task description. Pre-trained language model (PLM) is required to recognize and classify literal, metaphor, and syntactic anomaly. Selection preference violation theory (SPV) is used to detect whether there is a relationship violation between the target word and the context word of a sentence.

In metaphor detection tasks, previous studies generally choose to use selection preference violation (SPV) recognition methods. Wilks (1975, 1978) recognizes metaphors by identifying the re-

lationship between the target word and the context word. If the target word is not common in the context of its surrounding words, there is a relationship violation between the target word and the context word, i.e., there is a metaphorical expression. Mao et al. (2019) construct an end-to-end metaphor recognition model, introducing the language theory SPV to directly guide deep neural network (DNN) design for end-to-end sequential metaphor recognition. Unlike (Mao et al., 2019), Choi et al. (2021) combines SPV with the metaphor recognition process MIP to achieve automatic metaphor recognition. The SPV is a essentially mismatch phenomenon. Let us consider an example: in the sentence "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" is not capable of chewing, and "wires" are not capable of chewing. Consider another example "The girl comforts the clock.". The "girl" is alive, the "clock" is inanimate, and the inanimate "clock" is not the one that needs life. The inanimate "clock" is not a proper argument for the "comfort" of the needy, and this example is 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.

Chandola et al. (2009) define anomaly as patterns in data that do not conform to a well-defined notion of normal behavior. For anomaly detection, which involves the discovery of patterns in data that do not conform to the expected behavior, is an important problem that needs to be dealt with in various domains. In natural language processing, syntactic collocation anomaly are among the common types of anomaly (Lunsford and Lunsford, 2008). Syntactic interpretation elucidates that the syntactic representations of anomalous sentences are similar to well-constructed sentences; whereas, in semantic description, the syntactic representations of anomalous sentences are presented as missing or violate between words, which are ultimately replaced by semantic representations (Ivanova et al., 2012). Metaphors are 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 collocation anomaly as metaphors. However, no one has yet specifically linked metaphors to syntactic collocation anomaly. Since metaphors and syntactic collocation anomaly are commonplace in life, it becomes crucial to automatically investigate how to deal with them.

In this paper, syntactic anomaly detection is introduced into the metaphor detection task from the perspective of dealing with syntactic anomaly and metaphors(see figure 1 for description of tasks). To meet the needs of this task, we employ GPT-3 to generate some syntactic anomaly data and construct a dataset called LMA. Specifically, we designe a specific prompt and use this prompt to customize each syntactic anomaly type. Guided by this prompt, GPT-3 generates sentences containing our specified syntactic anomaly types. Our syntactic anomaly types are categorized into different fine-grained levels: 1) verb-noun anomaly, 2) adjective-noun anomaly, 3) adverb-verb anomaly, and 4) noun-verb anomaly. In our work, we mainly emphasize the effectiveness of sentence-level classification of metaphors and syntactic anomaly in a multi-task setting. This research aims to advance a deeper understanding of the correlation between syntactic anomaly and metaphors by introducing the syntactic anomaly dataset LMA, which provides a practical resource for multitask learning.

In summary, our contributions are as follows:

1. Firstly, we focus on the relationship between metaphor and syntactic anomaly, introducing syntactic anomaly detection as part of the metaphor detection task.

2. We successfully construct a dataset (LMA) that comprehensively contains literal, metaphor and syntactic anomaly. In this dataset, both metaphorical sentences as well as syntactically anomalous sentences account for 15%, while the rest are literal sentences.

3. 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

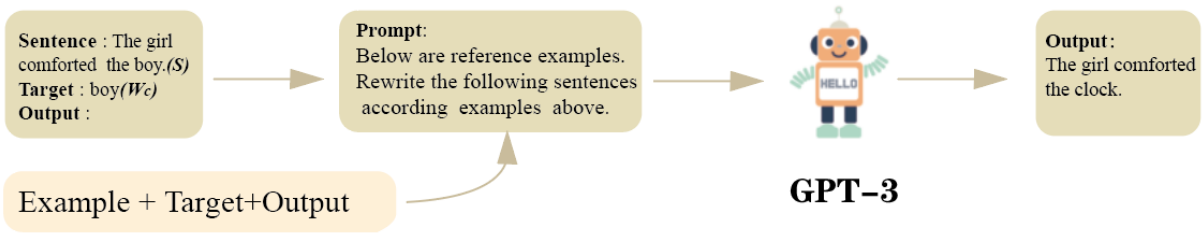

Figure 2: Flowchart of data generation. We take the example of generating verb-noun syntactic anomaly. Given a normal literal sentence ***S***, and the target word of the sentence ***Wc***. GPT-3 performs the same lexical modification of the target word of the input sentence under prompt.

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 metaphors, and Le et al. (2020) uses a textual dependency tree structure to construct metaphors. 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.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). Mancini et al. (2014), on the other hand, analyze syntactic anomaly of subject-verb inconsistency for person and number in Italian. Jia et al. (2018) attempts to construct a knowledge graph using subject-predicate-object ternary consistency relations, which in turn leads to the development of an anomaly detection system. Bock and Miller (1991) point out that speakers may commit subject-predicate agreement errors when

Below are some reference examples about anomalous type of Adjective-noun. Rewrite the sentence according examples about anomalous Adjective-noun of target. The "index" represents the index position of target.
Example 1: It was very difficult for my friends to call me with the small phone. **Target**: small (index: 12) **Output**: It was very difficult for my friends to call me with the delicious phone...
Example 2: You never want to make a man the centre of your existence. **Target**: your (index: 10) **Output**: You never want to make a man the centre of their existence.
Example 3: Manuals which may contain maps, schematic diagrams, and other materials warrant separate consideration. **Target**: schematic (index: 5) **Output**: Manuals which may contain maps, humble diagrams, and other materials warrant separate consideration.
Example 4: Early in the morning, the sunlight pours in the quiet garden. **Target**: quiet (index: 9) **Output**: Early in the morning, the sunlight pours in the delicate garden.
Sentence: Oh dear, Miss Williams said on an indrawn breath. **Target**: indrawn (index: 7) **Output**: [generated sentences]

Table 1: Hints for generating syntactic anomaly. We demonstrate this with adjective-noun anomaly, setting up four sets of examples to guide the model to generate syntactically anomalous sentences step by step.

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. Other work has studied adverb-verb morphological mismatches (Dickey et al., 2008; Nanousi et al., 2006; Stavrakaki and Kouvava, 2003; Tyler et al., 1990; Wenzlaff and Clahsen, 2004) (e.g. Tomorrow he walked). Dragoy et al. (2012) report a mismatch anomaly between the verb form and the time range in which the adverb was previously set (present adverb - past tense verb; past adverb - present tense verb). de Vega et al. (2010) explore the Spanish verb-adverb anomaly, where they propose that only verbs have temporal inflection suffixes, while adverbs convey temporal information through lexis rather than morphology. 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.

2.3 Large Language Model

Large Language Models (LLMs) are deep learning models that employ a huge number of parameters, typically ranging in size from billions to hundreds of billions of parameters. As a basis for the design of multiple LLMs, Transformer (Vaswani et al., 2017) introduces a self-attention mechanism to better capture the relationships between different locations in the input sequence . Based on Transformer, researchers carry out a study of the model BERT (Devlin et al., 2018), which represents a bidirectional encoder representation of Transformer . Liu et al. (2019) optimized the BERT especially. The final proposed RoBERTa is able to match or exceed all BERT methods in terms of performance (Liu et al., 2019). Recently, researchers (Lewis et al., 2019; Yoo et al., 2021) have attempted to explore new paradigms in the field of Natural Language Processing (NLP) using pre-trained models. GPT-3 is one of the largest language models to date, also using the Transformer architecture. Notably, LLMs are capable of learning with fewer sample prompts, and more and more research is beginning to focus on prompting mechanism-based approaches (Reynolds and McDonell, 2021; Schick and Schütze, 2020; Shin et al., 2020; Jiang et al., 2020; Zhao et al., 2021). We are witnessing another important shift in the NLP paradigm. To the best of our knowledge, this is the first task that proposes the use of a prompt-based approach to generate a specific syntactic anomaly dataset from a large language model, combined with metaphor recognition.

3 Method

3.1 Mission Description

In this paper, syntactic anomaly detection is introduced on top of the metaphor detection task. Metaphor is defined as a conceptual mapping between the source and target domains, where the target domain is interpreted through the source domain concepts (Lakoff and Johnson, 1980). Syntactic anomaly, on the other hand, refer to the phenomenon that the grammatical structures in a sentence do not conform to canonical linguistic rules, such as subject-predicate inconsistency (Ivanova et al., 2012). In the empirical study, we first design a prompt that guide GPT-3 to generate syntactically anomalous sentences through step-by-step prompts. The types of syntactic anomaly include adverb-verb anomaly, adjective-noun anomaly, verb-noun anomaly, and noun-verb anomaly, which are characterized by inconsistencies between words. For example, consider an original sentence, *"Let's go to the flicks."*. The syntactically anomalous sentences generated by GPT-3 may be shown below:

Example: The girl comforted the boy.
Target: boy(indxe:4)
Output: The girl comforted the clock.
Sentence: Let's go to the flicks.
Target: flicks(index:4)
Output: Let's go to the shoes.

Here we provide GPT-3 with an example of a syntactic anomaly modification, i.e., "This girl is comforting this boy.". Then, we provide the model with the target word and its index to generate relevant collocation exceptions. For example, if the target word is a noun, the possible exception type is a verb-noun exception. In this case, the generated sentence might be, *"Let's go to the shoes"*. Through this process, we construct a dataset containing lit-

eral meaning, different types of syntactic anomaly and metaphor, which provides strong support for further research on syntactic anomaly detection.

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) are generally studied at the sentence level as well. Sentence-level annotation methods usually involve categorizing entire sentences or text passages. The method first takes the entire text passage as input, marks the position of the words to be detected in the sentence, and then assigns a label or classification to the entire sentence. Our classification system consists of three classifications and nine classifications. Among them, the three classifications include metaphor, syntactic anomaly and literal meaning. The nine classifications, on the other hand, subdivide metaphors and syntactic anomaly according to lexical properties (adjective, noun, verb, and adverb) while considering literal meanings. We emphasize the classification of metaphor detection and syntactic anomaly detection at the sentence level.

3.2 Prompt Construction

In the tasks of this paper, prompt play a crucial role, especially in generating syntactic anomaly data. We design a prompt whose process consists of being given a set of prompt examples and then sampling from these examples using GPT-3. 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. We reference current literature based on GPT-3 prompts (Yoo et al., 2021; Reynolds and McDonell, 2021) in developing the prompt. Further, we adapt these generic templates to more closely match the tasks in this paper. For each syntactic anomaly type, we customize different task descriptions and examples to make them more relevant. Table 1 shows our specific prompts (in the case of Adjective-noun). Other types of anomalous sentences are generated in a similar way.

4 Dataset

This section delves into the construction of the syntactic anomaly dataset (LMA) that we construct. We are modifying and innovating on the basis of the VUA dataset variant.

4.1 Traditional Datasets

4.1.1 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 tags sentences using the MIPVU metaphor recognition program, which is guided by the principle of treating the literal meaning as the more basic or concrete meaning of a word.VUAMC is the largest publicly available annotated corpus of token-level metaphor detection, and the only one that studies the metaphorical nature of dummy words. The corpus contains 115 texts of four different types, covering academic, conversational, fictional, and journalistic texts. Based on VUAMC, VUA also derives a number of related variants.

4.1.2 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 adjectives, verbs without have, do, be, nouns and adjectives) in a sentence are labeled, while VUA Verb contains only verbs. However, in previous studies (Song et al., 2021; Feng and Ma, 2022; Wan et al., 2021; Su et al., 2020), VUA ALL POS was extended to include not only real-sense words but also dummy words, and contain a total of 205,425 samples, of which 116,622 were used for training, 38,628 for validation, and 50,175 for testing. In order 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.

4.2 Dataset Construction

4.2.1 Data Collection and Screening

We fully consider the key issues of quantity and distribution. First, we carefully screen a total of 25,760 sentences for initial screening by analyzing lexical labels (e.g., adverbs, verbs, adjectives, nouns). Among these sentences, the proportion of metaphorical samples is about 14.5%; the rest are categorized as non-metaphorical sentences. In order to better construct the dataset, we use some targeted strategies. Specifically, we extract sentences from the non-metaphorical samples according to a randomized step size that is comparable to the

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

Model	metaphor-literal			syntactic anomaly-literal			three-classification			nine-classification		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT-bs	0.838	0.842	0.841	0.838	0.842	0.841	0.754	0.759	0.756	0.736	0.758	0.740
BERT-lg	0.845	0.849	0.847	0.843	0.849	0.846	0.757	0.765	0.759	0.736	**0.762**	0.742
RoBERTa-bs	0.860	0.853	0.856	**0.863**	0.854	0.857	0.761	0.764	0.761	0.745	0.759	0.750
RoBERTa-lg	**0.876**	**0.873**	**0.874**	0.862	**0.872**	**0.859**	**0.771**	**0.779**	**0.772**	**0.773**	0.757	**0.761**

Table 2: Experiment 1 Results presentation. We conduct the evaluation on four baselines. The "**bs**" stands for the "base" version of the baseline model. The "**lg**" denotes the "large" version of the baseline model. Our experiments include metaphor-literal detection, syntactic anomaly detection, and three-classification and nine-classification detection. The evaluation metrics include precision (P), recall (R) and composite metric (F1), where F1 is the core metric.

proportion of metaphorical samples. This portion of data will be used to construct syntactic anomaly samples to improve the richness of the dataset.

4.2.2 GPT-3 Generation

We use as input to GPT-3 the literal sentences previously extract via randomized step size. Subsequently, following the prompt presented in the methodology of Section 3.2, we perform the modification and generation of syntactically anomalous sentences (see Figure 2).

4.2.3 Data Segmentation

Data segmentation consists of two steps: merging the data and dividing the dataset. We replace the syntactic anomaly samples generated by GPT-3 with the original samples to form the merged syntactic anomaly dataset (LMA). Subsequently, we divide the merged dataset into training, validation, and test sets, with division ratios of 0.7, 0.15, and 0.15, respectively. In the three-classification experiment, the training set contains 17,234 samples, the validation set contains 4,234 samples, and the test set contains 4,292 samples. While in the nine-classification experiment, the training set contains 17279 samples, the validation set contains 4206 samples and the test set contains 4275 samples.

5 Experiments

In this chapter, we describe our experimental design in detail. In Section 5.1, we review the traditional baseline approach. Then, in Section 5.2, we will provide relevant details of the experiment. Finally, we will discuss the hyperparameter tuning in the experiment as well as an in-depth analysis of the results.

5.1 Baseline

We conduct experiments on the following baseline model:

BERT:BERT (Devlin et al., 2018) employs a bidirectional Transformer encoder, available in both base and large versions. The model is able to consider all the words in the context at the same time, capturing the contextual information more comprehensively. In the pre-training phase, BERT performs two tasks: firstly, Masked Language Model (MLM), which randomly masks some words in the input text, and the model needs to predict these masked words; and secondly, Next Sentence Prediction (NSP), which the model needs to judge whether two sentences are adjacent in the text. Through fine-tuning, BERT can be adapted to different tasks and achieve excellent performance in natural language processing (NLP)

RoBERTa: (Liu et al., 2019) proposed an improved training scheme for the BERT model. Unlike BERT, RoBERTa removes the NSP task in pre-training, i.e., it no longer determines whether two sentences are adjacent. Meanwhile, RoBERTa uses larger scale training data and performs longer training steps to further improve the model performance.

5.2 Experimental Design

We conduct two sets of experiments in sentence-level annotation, each covering different fine-grained sub-experiments.

Experiment 1:The first set of experiments introduce BERT-base, BERT-large, RoBERTa-base and RoBERTa-large as baseline models. We focus on the classification performance of these models for anaphora and syntactic anomaly. The sub-experiments of Experiment 1 include two-

Model	metaphor-literal			syntactic anomaly-literal			three-classification		
	P	R	F1	P	R	F1	P	R	F1
BERT-bs*	0.848	0.839	0.843	0.827	0.844	0.843	0.677	0.684	0.678
BERT-lg*	0.855	0.864	0.857	0.844	0.850	0.847	0.685	0.702	0.689
RoBERTa-bs*	0.856	0.859	0.857	0.853	**0.861**	0.856	**0.726**	0.704	0.711
RoBERTa-lg*	**0.876**	**0.880**	**0.878**	**0.866**	0.856	**0.861**	0.721	**0.740**	**0.729**

Table 3: Experiment 2 results are presented. We introduce SPV and measure it on four baselines. The experiments include metaphor-literal detection, syntactic anomaly-literal detection, and three-classification detection. The metrics are the same as in Experiment 1. "*" stands for "SPV". bs stands for the base version of the baseline model. lg stands for the large version of the baseline model.

classification detection, three-classification detection and nine-classification detection. Among them, the two-classification detection includes metaphor-literal detection, syntactic anomaly-literal detection, and metaphor-syntactic anomaly detection. Three-classification detection includes literal-metaphor-syntactic anomaly detection. Nine-classification detection is a further subdivision of metaphorical and syntactic anomaly according to lexical labels (adjective, noun, verb, adverb) based on three-classification detection. Two-classification detection is designed as a controlled experiment. The dataset is divided according to the rules in Section 4.2.3.

Experiment 2:In the second set of experiments, we introduce SPV, i.e., combining SPV with BERT-base, BERT-large, RoBERTa-base, and RoBERTa-large 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 syntactic anomaly and metaphor classification tasks. The sub-experiments of Experiment 2 mainly consist of dichotomous and trichotomous detection. Among them, the categories of dichotomous and tricotomous are the same as in Experiment 1.

6 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.

7 Experimental Results

7.1 Evaluation Metric

In our sentence-level task, we consider three widely used evaluation metrics, namely precision, recall, and F1 score. These metrics provide a comprehensive assessment of the model's performance. Precision measures the extent to which the model correctly predicts, focusing on the proportion of samples that the model determines to be in the positive classification that are actually in the positive classification. Recall measures the ability of the model to correctly identify samples in the positive classification (true instances). The F1 score is a combination of precision and recall metrics is used to balance the accuracy and recall of the model. By using these three metrics, we are able to comprehensively evaluate the performance of the model in sentence-level tasks, providing an assessment of the model in different aspects of the task and enabling a more complete understanding of the model's effectiveness in the task.

7.2 Results and Analysis

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

In Experiment 1, we evaluate the performance of BERT-base, BERT-large, RoBERTa-base and RoBERTa-large on literal, metaphor and syntactic anomaly classification tasks, and the results are shown in Table 2. Observing the two sub-experiment index scores of metaphor-literal detection and anomaly-literal detection, we can find

Model	metaphor-syntactic anomaly		
	P	R	F1
BERT-bs	0.803	0.802	0.802
BERT-lg	0.817	0.814	0.813
RoBERTa-bs	**0.845**	0.844	0.844
RoBERTa-lg	0.844	**0.846**	**0.846**
BERT-bs*	0.719	0.717	0.717
BERT-lg*	0.736	0.735	0.735
RoBERTa-bs*	0.738	0.735	0.736
RoBERTa-lg*	**0.778**	**0.775**	**0.777**

Table 4: Experiment 2 results are presented. The assessment tasks includ metaphor-syntactic anomaly detection. The metrics are the same as in Experiment 1. "*" stands for "SPV". The "**bs**" stands for the "base" version of the baseline model. The "**lg**" denotes the "large" version of the baseline model.

that the models not only perform well in metaphor recognition, but also achieve better performance in syntactic anomaly recognition. RoBERTa-large achieves F1 scores of 0.874 and 0.859 on these two tasks respectively. In the three-classification sub-experiment, the performance of all four baseline models decreased, with RoBERTa-large achieving the highest F1 score of 0.772 (decreasing by 0.102 and 0.087, respectively). Looking further at the results of the nine-classification experiments, we can see that the performance of the baseline model further decreases compared to the three-classification experiments. This may be due to the fact that the model has to further differentiate lexical label types for metaphors and syntactic anomaly in the nine-classification sub-experiment, which leads to an increase in the difficulty of classification.

In Experiment 2, we explore the effect of SPV on metaphor and syntactic anomaly recognition, the results of which are shown in Table 3 and 4. Comparing the F1 scores for metaphor-literal detection as well as syntactic anomaly literal detection in Table 2 and table 3, we can see that the performance of all four baseline models has increased, with RoBERTa-large achieving F1 scores of 0.878 (an improvement of 0.004) and 0.861 (an improvement of 0.002), respectively, as compared to Experiment 1. In the triple classification sub-experiment, we note that instead of an increase in the model's performance, there is a decrease, with RoBERTa-large's F1 score dropping to 0.729 (a decrease of 0.044). Further we observe in Table 4 that for metaphor-syntactic anomaly detection, the four baseline models have reduced scores after the introduction of SPV. The results of Experiment 2 suggest that the introduction of SPV in multitasking scenarios may lead to complex cross-influences that exacerbate the confusion of the models for metaphorical and syntactic anomaly.

8 Conclusion

In metaphor detection tasks, collocations are not only metaphors, but may also be syntactic collocation anomaly. This study focuses on the analysis of metaphor and syntactic anomaly at different levels of granularity. We design a specialized prompt for this task, based on which we call GPT-3 to generate data for four syntactic collocation anomaly. Using the syntactic anomaly data, we further construct a dataset LMA containing literals, metaphors, and syntactic anomaly. In metaphor detection, the theory of selection preference violation (SPV) is commonly used. We also explore the role of SPV for metaphor and syntactic anomaly detection. 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. Accurate identification of metaphorical and syntactic anomaly is crucial. We hope that this study will help related researchers to better distinguish syntactic anomaly.

9 Limitations

In this study, we propose a task that specializes in anaphora and syntactic anomaly. We use GPT-3 in constructing the anomaly data. Despite the high performance of GPT-3, there are some discrepancies. It is possible that every piece of data generated GPT-3 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 metaphors and syntactic anomaly.

10 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.

11 Acknowledgements

I would like to express my sincere gratitude to my supervisors and mentors in the laboratory during the completion of this thesis. Their professional guidance and attentive nurturing had a profound impact on my academic growth. My mentor and senior brothers gave me critical support and guidance throughout the research process, providing me with valuable academic resources and insights.

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.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

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*.

Manuel de Vega, Mabel Urrutia, and Alberto Dominguez. 2010. Tracking lexical and syntactic processes of verb morphology with erp. *Journal of Neurolinguistics*, 23(4):400–415.

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*.

Michael Walsh Dickey, Lisa H Milman, and Cynthia K Thompson. 2008. Judgment of functional morphology in agrammatic aphasia. *Journal of Neurolinguistics*, 21(1):35–65.

Olga Dragoy, Laurie A Stowe, Laura S Bos, and Roelien Bastiaanse. 2012. From time to time: Processing time reference violations in dutch. *Journal of Memory and Language*, 66(1):307–325.

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.

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

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

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.

Bin Jia, Cailing Dong, Zhijiang Chen, Kuo-Chu Chang, Nichole Sullivan, and Genshe Chen. 2018. Pattern discovery and anomaly detection via knowledge graph. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 2392–2399. IEEE.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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*.

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*.

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.

Simona Mancini, Francesca Postiglione, Alessandro Laudanna, and Luigi Rizzi. 2014. On the person-number distinction: Subject-verb agreement processing in italian. *Lingua*, 146:28–38.

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.

Vicky Nanousi, Jackie Masterson, Judit Druks, and Martin Atkinson. 2006. Interpretable vs. uninterpretable features: Evidence from six greek-speaking agrammatic patients. *Journal of Neurolinguistics*, 19(3):209–238.

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.

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*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

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.

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.

Stavroula Stavrakaki and Sofia Kouvava. 2003. Functional categories in agrammatism: Evidence from greek. *Brain and Language*, 86(1):129–141.

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*.

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.

Lorraine K Tyler, Susan Behrens, Howard Cobb, and William Marslen-Wilson. 1990. Processing distinctions between stems and affixes: Evidence from a non-fluent aphasic patient. *Cognition*, 36(2):129–153.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

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.

Michaela Wenzlaff and Harald Clahsen. 2004. Tense and agreement in german agrammatism. *Brain and language*, 89(1):57–68.

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. 2023. Adversarial multi-task learning for end-to-end metaphor detection. *arXiv preprint arXiv:2305.16638*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.