Multilingual Multi-Figurative Language Detection

Huiyuan Lai, Antonio Toral, Malvina Nissim
CLCG, University of Groningen / The Netherlands
{h.lai, a.toral.ruiz, m.nissim}@rug.nl

Abstract

Figures of speech help people express abstract concepts and evoke stronger emotions than literal expressions, thereby making texts more creative and engaging. Due to its pervasive and fundamental character, figurative language understanding has been addressed in Natural Language Processing, but it's highly understudied in a multilingual setting and when considering more than one figure of speech at the same time. To bridge this gap, we introduce *multilingual multi-figurative language modelling*, and provide a benchmark for sentence-level figurative language detection, covering three common figures of speech and seven languages. Specifically, we develop a framework for figurative language detection based on template-based prompt learning. In so doing, we unify multiple detection tasks that are interrelated across multiple figures of speech and languages, without requiring task- or language-specific modules. Experimental results show that our framework outperforms several strong baselines and may serve as a blueprint for the joint modelling of other interrelated tasks.

1 Introduction

Figurative language is ubiquitous in human language, allows us to convey abstract concepts and emotions, and has been embedded intimately in our cultures and behaviors (Roberts and Kreuz, 1994; Harmon, 2015). In the hyperbolic sentence "*My heart failed a few times while waiting for the result.*", the expression "*my heart failed a few times*" is not a literal heart-stop, it exaggerates the mood of when waiting for a possibly important result, thereby vividly showing anxiety.

Recent years have seen a lot of interest in figurative language processing in the NLP community, including the successful organization of dedicated workshops (Beigman Klebanov et al., 2018; Klebanov et al., 2020; Ghosh et al., 2022). There are many works focusing on figurative language detection, mostly in English, including hyperbole (Troiano et al., 2018), metonymy (Nissim and Markert, 2003), metaphor (Tsvetkov et al., 2014), idiom (Liu and Hwa, 2018) and sarcasm (Hazarika et al., 2018). In addition, researchers have started start to pay attention to figurative language detection in a multilingual scenario (Tsvetkov et al., 2014; Tedeschi et al., 2022; Aghazadeh et al., 2022; Tayyar Madabushi et al., 2022), where models can exploit cross-lingual knowledge transfer (Conneau and Lample, 2019). Nonetheless, detection tasks for different figures of speech are usually studied independently of each other, which leads to having to train separate models for each figure of speech. However, different figures of speech are often related to each other, and therefore models can thus potentially benefit from cross-figurative knowledge transfer, as empirically shown by Lai and Nissim (2022) in a monolingual setting for English.

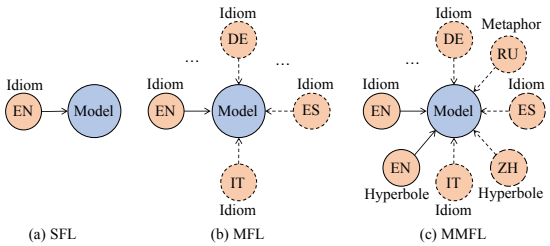


Figure 1: Overview of three different modelling scenarios: (a) single figurative language (SFL), (b) multilingual figurative language (MFL) and (c) multilingual multi-figurative language (MMFL).

In this paper we investigate how these related detection tasks can be connected and modelled jointly in a multilingual way (see Table 1). To do so, we propose a multitask framework to model multilingual multi-figurative language detection at the sentence level. As shown in Figure 1, our goal is to connect the detection tasks from different languages and different figures of speech, resulting in a unified model which can benefit from cross-

Findings of the Association for Computational Linguistics: ACL 2023, pages 9254–9267
July 9-14, 2023 ©2023 Association for Computational Linguistics

Reference	Description	M-Lang	M-Fig	M-Task	Level
Troiano et al. (2018)	Hyperbole detection in English	✗	✗	✗	Sentence
Tedeschi et al. (2022)	Multilingual idiom detection	✓	✗	✗	Word
Tayyar Madabushi et al. (2022)	Multilingual idiom detection	✓	✗	✗	Sentence
Aghazadeh et al. (2022)	Multilingual metaphor detection	✓	✗	✗	Sentence
Lai and Nissim (2022)	Multi-figurative language generation in English	✗	✓	✓	Sentence
Our work	Multilingual multi-figurative language detection	✓	✓	✓	Sentence

Table 1: A comparison of different previous works and the present one, according to whether they perform word-/sentence-level detection in a multilingual (M-Lang), multi-figurative (M-Fig), or multi-task (M-Task) fashion.

lingual and cross-figurative knowledge transfer.

Generally, a multi-task framework consists of shared modules and task-specific modules. With the development of pre-trained language models (PLMs), prompt learning offers the opportunity to model multiple tasks in a framework that does not require task-dependent parameters (Radford et al., 2019; Brown et al., 2020; Fu et al., 2022; Mishra et al., 2022). With this method, task-specific language instructions are predefined and used to guide the model to handle different tasks.

In practice, we first formalize the figurative language detection task as a text-to-text generation problem, where the input is the source sentence while the target is a textual label (e.g. "literal" or "idiomatic"). This method thus enables us to train our models in a sequence-to-sequence (seq2seq) fashion. We then prepend the prompt template to source sentences from various tasks when feeding them into the model. This connects multiple figures of speech and languages in a unified framework, also leading to a better understanding of how to jointly model tasks related to each other. We perform extensive experiments on three figures of speech: hyperbole, idiom, and metaphor, involving seven languages (English:EN, Chinese:ZH, German:DE, Spanish:ES, Italian:IT, Farsi:FA, and Russian:RU).

Our main contributions are as follows: (i) We introduce the novel task of multilingual multi-figurative language detection, wherein we explore the potential of joint modelling. (ii) We introduce a multitask and multilingual framework based on prompt learning, which unifies interrelated detection tasks without task- nor language-specific modules. (iii) We evaluate the model's generalization capabilities across a range of figures of speech and languages: extensive experiments are run for different settings, including in-language, zero-shot, cross-lingual, and cross-figurative to show how the unified framework performs. (iv) Our framework may serve as a blueprint for joint modelling of other interrelated tasks, such as the detection of hate speech (Waseem and Hovy, 2016), offensive and abusive language (Caselli et al., 2020), toxicity (Pavlopoulos et al., 2020), as well as fake news and AI-generated content (Zellers et al., 2019). We have released our code and all preprocessed dataset.[1]

2 Related Work

We briefly introduce the background of figurative language detection, from feature engineering to neural-based approaches, as well as prompt-based learning with PLMs.

2.1 Figurative Language Detection

This task often involves word-level and sentence-level detection. Word-level detection is concerned with identifying the exact words within the context of a sentence used with a figurative meaning. Sentence-level detection, as a binary classification problem, requires to automatically detect whether a given sentence is literal or not.

Feature Engineering Traditionally, researchers have investigated hand-engineered features to understand figurative usages. These features are primarily concerned with linguistic aspects, including word imageability (Broadwell et al., 2013; Troiano et al., 2018), word unexpectedness (Troiano et al., 2018), syntactic head-modifier relations (Nissim and Markert, 2003), abstractness and semantic supersenses (Tsvetkov et al., 2014), property norms (Bulat et al., 2017), pragmatic phenomena (Karoui et al., 2017), together with other aspects such as sentiment (Troiano et al., 2018; Rohanian et al., 2018) and sensoriality (Tekiroğlu et al., 2015). These features rely heavily on manual extraction and are very much task-dependent. Exploiting verb and noun clustering (Shutova et al.,

[1] https://github.com/laihuiyuan/MMFLD

2010) and bag-of-words approaches (Köper and Schulte im Walde, 2016) are common automated methods to reduce manual work.

Neural-Based Approaches In the last decade, researchers have moved from feature engineering to neural-based modelling, using LSTM- (Wu et al., 2018; Gao et al., 2018; Mao et al., 2019; Kong et al., 2020) and CNN-based approaches (Wu et al., 2018; Kong et al., 2020) for figurative language detection. Most recently, PLMs have been used for this task, usually yielding new state-of-the-art results (Su et al., 2020; Choi et al., 2021; Zeng and Bhat, 2021; Tedeschi et al., 2022). Similar to other NLP tasks, researchers have also moved towards multilingual detection (Tsvetkov et al., 2014; Tedeschi et al., 2022; Tayyar Madabushi et al., 2022; Aghazadeh et al., 2022), especially thanks to cross-lingual knowledge transfer via multilingual PLMs. All these works focus on single figures of speech, i.e. detecting whether a sentence (or each word in a sentence) contains a given figure of speech or it is literal. We take here the first step towards multilingual multi-figurative language modelling to introduce a unified framework for multiple languages and multiple figures of speech, focusing on sentence-level detection.

2.2 Pre-Training and Prompt Learning

Over the past few years, PLMs have brought NLP to a new era (Devlin et al., 2019; Radford et al., 2018, 2019; Brown et al., 2020). PLMs are pre-trained on massive textual data in a self-supervised manner, and then fine-tuned on downstream tasks with task-specific training objectives. This paradigm, however, has to be adapted to different target tasks, where the task-specific objectives are different from the pre-training one, and the introduction of additional parameters such as a PLM-based classifier is at times necessary.

Prompt learning, a new learning paradigm based on PLMs, aims to make better use of pre-trained knowledge by reformulating tasks to be close to the pre-training objectives (Liu et al., 2022). Specifically, this is a method of leveraging PLMs by prepending task-specific prompts to the original input when feeding it into PLMs. One way to do this is with manually designed templates as task instructions (Radford et al., 2019; Raffel et al., 2020); another one is to use continuous prompts that optimize a sequence of continuous task-specific vectors (Lester et al., 2021; Li and Liang, 2021). More recently, Fu et al. (2022) have introduced an mT5-based framework to learn a unified semantic space blurring the boundaries of 6 NLP tasks with the prompting method, which we adopt in this work. Here, we investigate how a small PLM such as mt5 can be used in the multilingual multitask prompting framework, also to better understand how interrelated tasks can benefit from such a scheme.

Compared to very large models like GPT-3 (Brown et al., 2020), smaller models have the significant advantage of lower hardware requirements, making it easier to customize them quickly and cheaply for specific tasks, to implement modelling ideas iteratively, and for other researchers to reproduce experiments, too. Using a small PLM could however be very challenging when modelling more unrelated NLP tasks than those addressed in previous and in the current work, so this is something to bear in mind for future extensions.

Form	Lang	Train	Valid	Test
Hyperbole	EN	3,352	100	300
	ZH	3,760	600	1,000
Idiom	EN	18,676	1,470	200 [41/159]
	DE	14,952	1,670	200 [19/181]
	ES	12,238	1,706	199 [66/133]
	IT	15,804	1,732	200 [48/152]
Metaphor	EN	12,238	4,014	4,014
	ES	12,238	2,236	4,474
	FA	12,238	1,802	3,604
	RU	12,238	1,748	3,498

Table 2: Dataset Statistics. The label distribution is completely balanced (50%/50%), except for the idiom test sets, where the distribution is indicated in brackets as the proportion of literal/idiomatic next to the totals.

3 Tasks and Datasets

3.1 Task Formulation

We focus on figurative language detection at sentence-level, which can be viewed as a binary classification task that requires identifying whether a given sentence is literal or non-literal (e.g. idiomatic). To unify multiple figurative language detection tasks in different languages, we reformulate them as a text-to-text generation problem, where our model will generate the textual label for each given sentence. For instance, given a sample s from a detection task $T_{idiomatic} \in T$, where $T = \{T_{hyperbole}, T_{idiom}, T_{metaphor}\}$ is the task set we consider, the model aims to output the text label $y \in \{\text{Literal}, \text{Idiomatic}\}$.

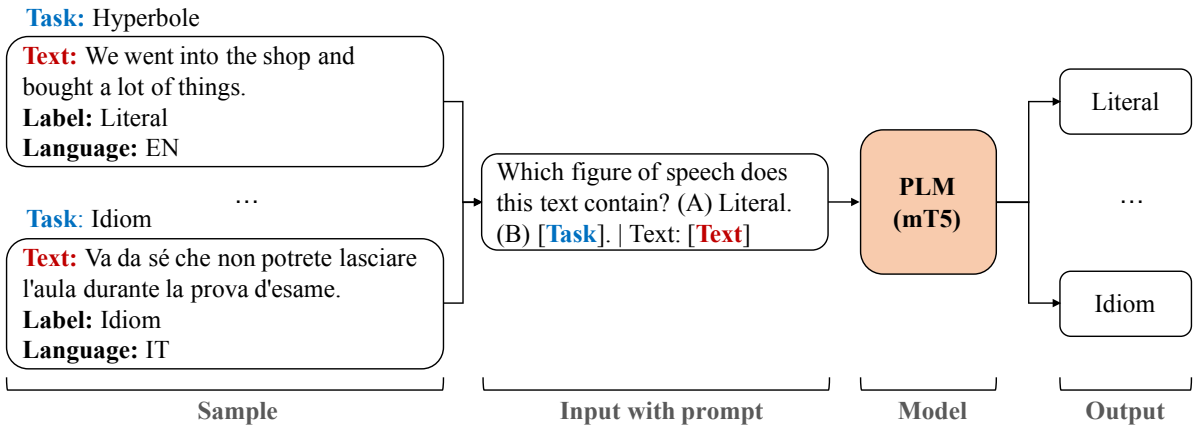


Figure 2: Overview for multilingual multi-figurative modelling based on prompt training. Given a detection task and the input text, we combine it with the predefined template, thus instructing the model to handle it.

3.2 Datasets

We use five existing figurative language datasets for our experiments, which cover three figures of speech and seven languages. Table 2 shows the dataset statistics for the various languages in each figure of speech.

Hyperbole HYPO (Troiano et al., 2018) is an English dataset containing 709 hyperbolic sentences with their corresponding non-hyperbolic versions. HYPO-Red (Tian et al., 2021) is another dataset that includes literal and hyperbolic texts. We combine these two datasets for the English hyperbole detection task. HYPO-cn (Kong et al., 2020) is a Chinese hyperbole detection dataset. Since both English and Chinese hyperbole datasets are rather small compared to the sizes of the training datasets for the other figures of speech, we upsample them by random instance replication obtaining training sets of 10,000 samples.

Idiom ID10M (Tedeschi et al., 2022) is a multilingual idiom dataset, containing automatically-created (silver) training and validation data in 10 languages and manually-created (gold) test sets in 4 languages: English, German, Italian, Spanish. This dataset is designed for word-level idiom detection; we convert it to sentence-level labels and use the four languages with gold data.

Metaphor LCC (Mohler et al., 2016) is a multilingual metaphor dataset derived from web-crawled data in four languages: English, Spanish, Russian, and Farsi. It provides metaphoricity ratings for within-sentence word pairs on a four-point scale, including 0 as no, 1 as weak, 2 as Conventional, and 3 as clear metaphor. We use the data preprocessed by Aghazadeh et al. (2022).

4 Multilingual Multi-Figurative Model

We propose a multitask and multilingual framework based on template-based prompt learning for figurative language detection.

4.1 Multitask Prompt Training

We use mT5 (Xue et al., 2021) as our backbone and jointly model multiple detection tasks, with the ultimate goal of having one single model that can handle the detection of multiple figures of speech in multiple languages. The overall framework is illustrated in Figure 2. Given a sample x from the t^{th} task T_t, it is first combined with the predefined prompt template p_t and then fed into model M, which is expected to produce the label y: $M(x, p_t) = y^{'}$.

We minimize the negative log-likelihood of the sequences of the model's outputs, the loss function being formulated as:

$$L_\theta = -\sum \log(\boldsymbol{y} \mid \boldsymbol{x}; \theta) \quad (1)$$

where θ are the parameters of mT5. $\boldsymbol{x}$ and $\boldsymbol{y}$ represents the sequences of the given sentence x and its text label y, respectively. We use the multilingual and multi-figurative samples from dataset T to fine-tune mT5, adapting it to the figurative language detection tasks.

We design the prompt templates based on our intuition of how we would ask a human annotator to complete the figurative language detection task. In our main framework, we use a cross-lingual

template setting whose templates for all tasks are in English. We will assess the impact of different prompts settings, including template and language (see Sec 5.4).

4.2 Generalization

We investigate the generalization ability of our proposed framework in cross-figurative and cross-lingual scenarios, where the training and test data come from different figure-of-speech or different languages.

Cross-Figurative Knowledge Inspired by Lai and Nissim (2022), we evaluate our framework in terms of cross-figurative knowledge transfer. The hypothesis is that different figures of speech might share some figurative features, and that a text may contain different figures of speech simultaneously, possibly triggered by different textual portions, so that a single framework might warrant a large knowledge gain through transfer from one figure of speech to another.

Using a multitask framework to jointly model multiple figurative language detection tasks, the cross-figurative generalization ability is expected to improve performance across different tasks compared to single figurative language modelling.

Cross-Lingual Knowledge Multilingual PLMs are pre-trained on texts from multiple languages in a self-supervised way, which enables different languages to be represented in a single space. Therefore, words and phrases that are similar across languages will be close to each other. We extend and evaluate the cross-lingual generalization in metaphor carried out by Aghazadeh et al. (2022) to a setting with multiple figures of speech. The hypothesis is that if the knowledge of figurative language is transferable across languages, then the model M_{l_m} would be able to have a good generalization in language l_n based on what it has learned in language l_m: $M_{l_m}(x, p_t) = y^{'}$, for $(x, y) \in T_t$ in language l_n. Furthermore, cross-lingual knowledge transfer can further improve model performance when doing multilingual modelling.

However, cultural differences often have a great influence on the usage of figurative language. Idioms are culture-/language-specific, for example, with established meanings over a long period of usage in a specific cultural background (Nunberg et al., 1994). Therefore, we expect that the model will have different performances in cross-lingual generalization for different figures of speech depending on how culturally-related the languages involved are.

5 Experiments

5.1 Setup

We use mT5-base (580M parameters) to evaluate our framework. All experiments are implemented atop Transformers (Wolf et al., 2020). We train our models with batch size 32, using the Adam optimiser (Kingma and Ba, 2015) with a polynomial learning rate decay. We set a linear warmup of 1,000 steps for a maximum learning rate of 1e-4 and a maximum decay of 10,000 steps for a minimum learning rate of 5e-5. We evaluate checkpoints every 1,000 steps, and use early stopping (patience 5) if validation performance does not improve. Following Aghazadeh et al. (2022), we report performance as detection accuracy for all experiments. In Section 5.4, we include an additional analysis for the unbalanced idiom datasets.

5.2 Model Settings

Since we take the first step towards the joint modelling for multilingual multi-figurative language detection, we conduct extensive experiments with different architectures and settings, leading to five sets of models. Additionally, we obtain zero-shot results in non-English languages by utilizing English-only variants of the same sets of models.

- **Baseline** Following Tayyar Madabushi et al. (2022), we train a binary detection classifier for each figure of speech and language by fine-tuning multilingual BERT (Devlin et al., 2019). Our work is similar to one previous work on sentence-level metaphor detection, which is carried out by Aghazadeh et al. (2022) in a multilingual setting. However, they assume that the phrase in the sentence to be classified as metaphoric or not is already known in advance, while our models do not use such information. Therefore, we do not consider it as a baseline model.
- **Vanilla mT5** Similar to mBERT, we fine-tune mT5 on specific figures of speech for each language but in a seq2seq fashion.
- **Prompt mT5** We fine-tune mT5 with the prompt template in a seq2seq way for each figure of speech in each language with the aforementioned sets of models.

Model	Hyperbole		Idiom				Metaphor			
	EN	ZH	EN	DE	ES	IT	EN	ES	FA	RU
Main results										
Baseline	72.33	80.40	79.00	72.50	66.33	70.50	81.37	80.11	74.83	79.93
Vanilla mT5	72.67	71.40	79.50	74.50	64.82	**76.00**	82.64	82.32	77.33	82.25
+ multitask	72.67	81.40	62.00	74.50	56.78	72.00	81.86	81.20	77.61	**83.76**
Prompt mT5	81.00	81.60	79.50	75.00	**68.34**	75.00	**83.43**	82.66	76.64	83.39
+ multitask	**82.00**	**82.60**	**86.00**	**79.00**	67.84	**76.00**	83.06	**83.10**	**78.14**	83.16
(Zero-shot) with EN model										
Baseline	72.33	69.60	79.00	62.00	61.81	60.00	81.37	71.70	61.29	69.01
Vanilla mT5	72.67	70.20	79.50	53.00	64.32	70.50	82.64	75.10	68.70	76.10
+ multitask	65.67	64.90	72.50	52.50	37.69	63.50	82.41	71.86	66.84	73.61
Prompt mT5	81.00	74.00	79.50	59.00	**69.85**	76.50	**83.43**	**75.95**	**70.17**	**76.39**
+ multitask	**82.33**	**76.10**	**81.50**	**65.60**	66.83	**79.50**	81.27	74.99	68.70	75.93

Table 3: Results (accuracy) for multilingual multi-figurative language detection, covering three figures of speech and seven languages. Notes: (i) we include results on English tasks for the block of zero-shot modelling with the EN model for comparison with those included in the main results; (ii) bold numbers indicate the best systems for each block, and underlined numbers indicate the best score for each language.

- **+ multitask** These are multilingual multitask models. We fine-tune mT5-based models with their corresponding single-task training methods using all data from *T*.
- **Zero-shot with EN model** Based on the above models, but we train them on English data only and test them on non-English languages.

5.3 Results

Table 3 reports results on three figurative language detection tasks in seven languages.

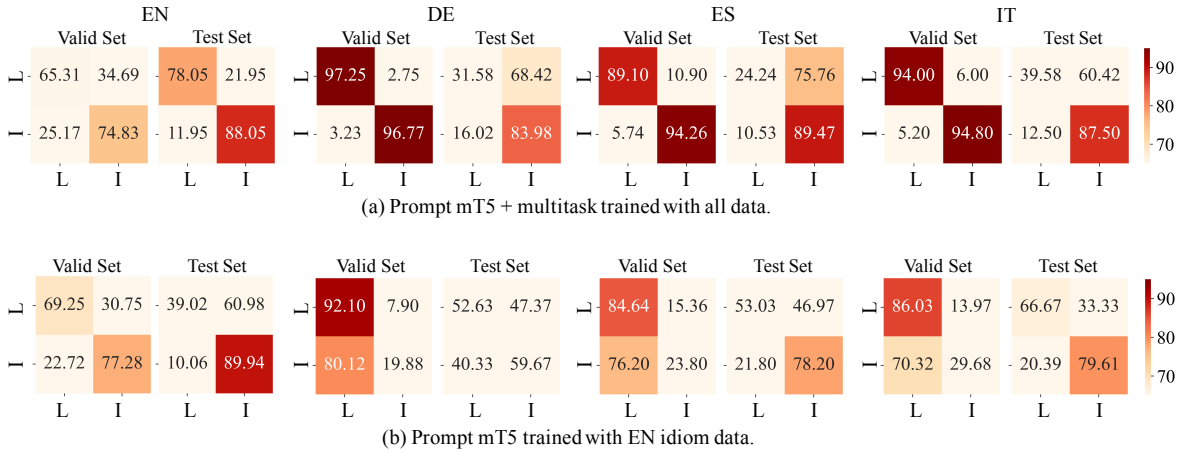
Main Results We see that `Vanila mT5` performs better than `mBERT` on most tasks, except ZH hyperbole and ES idiom. When `Vanila mT5` is used for multitask training, unsurprisingly, its performance drops in many tasks. One straight reason is that it is challenging to model multiple tasks at once. The other possible reason is that a text may contain features of multiple figures of speech at the same time, but there is not enough evidence to guide the model to perform a specific task. In other words, the model may correctly predict the figurative form for a given text, but it does not match the label of the target task.

When looking at `Prompt mT5`, we see that the model with prompt training brings improvement for most tasks compared to `Vanila mT5`. This shows the effectiveness of the prompt, which instructs the model to perform the target task. `Prompt mT5` with multi-task training has the best performances on most tasks: (i) it shows a steady improvement in hyperbole detection; (ii) in idiom detection performances are boosted for EN, DE, and IT though the ES score is lower compared to `Prompt mT5`; (iii) for metaphor detection it achieves the highest accuracy in ES and FA but slightly underperforms in EN and RU compared to `Prompt mT5`.

Zero-Shot For zero-shot results on non-EN languages using EN models, we see similar trends to the main results (see Table 3, second block). `Vanilla mT5` has overall better performances than its multitask counterpart and `mBERT`. We observe that `Prompt mT5`-based models have a clear edge in this setting, with the highest accuracy for all tasks and languages obtained by one of them. EN models yield the highest accuracy scores in EN hyperbole and metaphor detection, and even in idiom detection of ES and IT with zero-shot. The main reason for this is most likely that the idiom training and validation data is created automatically, leading to a non-test set of inferior quality and reduced performance on the test set compared to the validation set (see Sec 5.4). Overall, a zero-shot approach for figurative language detection when lacking high-quality resources in the target language seems a highly reliable strategy.

5.4 Analysis and Dissusion

Error Analysis. Table 4 presents the results of our main model (`Prompt mT5` + multitask) on validation and test sets. The performances on the test sets are comparable to the validation sets for hyperbole and metaphor, while the idiom task stands out: for EN idiom detection, test accuracy is higher than validation accuracy, while we observe the opposite



(a) Prompt mT5 + multitask trained with all data.

(b) Prompt mT5 trained with EN idiom data.

Figure 3: Confusion matrix of main model and zero-shot on idiom detection, where predictions (X-axis) are compared to the corresponding ground truth labels (Y-axis). L = Literal and I = Idiomatic.

Form	Lang	Valid	Test	Lang	Valid	Test
Hyperbole	EN	87.00	82.00	ZH	83.00	82.60
Idiom	EN	70.07	86.00	DE	97.01	79.00
	ES	91.68	67.84	IT	94.40	76.00
Metaphor	EN	83.06	83.06	ES	83.54	83.10
	FA	78.30	78.14	RU	82.78	83.16

Table 4: Results (accuracy) of our main model (`Prompt mT5 + multitask`) on validation and test sets.

Lang	Valid		Test	
	Literal	Idiomatic	Literal	Idiomatic
EN	34.83	49.12	48.78	62.26
DE	2.16	97.61	63.16	65.75
ES	18.17	97.19	13.64	30.83
IT	5.88	98.28	85.42	68.42

Table 5: The ratio (%) of idiomatic expressions contained in sentences of valid/test sets that appear in idiomatic sentences of training sets.

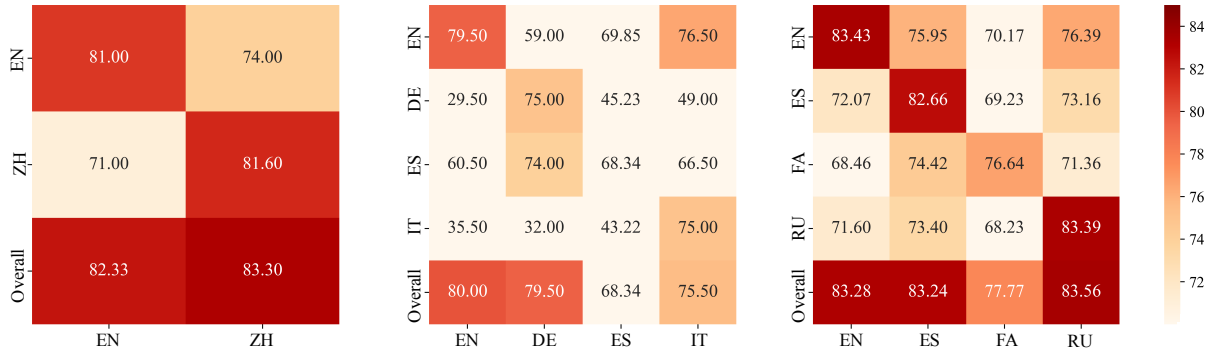
in languages other than English, with validation scores above 90% and test scores below 80%.

To analyze this behaviour of the idiom task, in Table 5 we report the ratio of idiomatic expressions contained in sentences of the validation/test sets that also appear in idiomatic sentences of the training sets. The distribution of the EN data is relatively balanced for both validation and test. For other languages, most of the expressions in the idiomatic sentences of the validation set are already present in the training set, but this is not the case for the literal sentences. Regarding test sets, the ratio of DE and IT are very high for both literal and idiomatic sentences, but very low for ES, which poses a significant challenge to the model.

We also group both the predictions and the labels to produce the confusion matrices for the main and EN models in Figure 3. For the EN task, we see that scores on the main diagonal are higher than those on the secondary diagonal, except for the EN model in the test set (39.02 vs 60.98). This is commonly observed in binary classification experiments. We have different observations on in-language training and zero-shot in other languages: (i) from (a), we see that the in-language model performs very well on the validation set for literal and idiomatic sentences, while it overpredicts literal sentences on the test set; (ii) in contrast, EN models in (b) perform better in the test sets and they overpredict idiomatic sentences on the validation set.

Generally, a sentence might not be idiomatic as a whole although it contains idioms. Such sentences will be labelled as non-literal in the automatic dataset creation. Based on the above observations, we see that the distribution of the automatically created training and validation data is quite different from the manually created test set, and the quality of the former is much lower than that of the latter. The nature of training data actually affects the stability of the model on different tasks. For instance, this even leads to better performances using EN models (zero-shot) on ES and IT than using in-language-trained models (see Table 3).

Cross-Figurative Knowledge Transfer To further investigate cross-figurative knowledge transfer, we sample different figures of speech for two languages from our dataset and compare single- to multi-figurative language modelling. Table 6 shows the results for EN and ES. For EN, com-

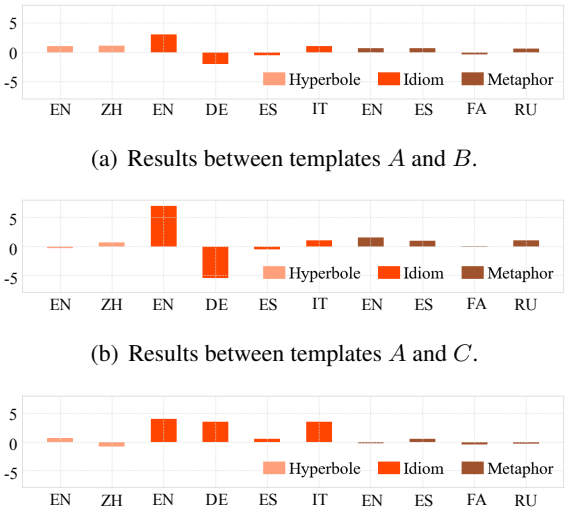


(a) Hyperbole (b) Idiom (c) Metaphor

Figure 4: Cross-lingual detection results (accuracy). The y-axis shows the language used for training (overall stands for a model trained on all the languages) while the x-axis indicates the language in which the model is tested on.

Model	Lang	Hyperbole	Idiom	Metaphor
Prompt mT5	EN	81.00	79.50	83.43
+ multitask		82.33	81.50	81.27
Prompt mT5	ES	-	68.34	82.66
+ multitask		-	70.35	82.14

Table 6: Results (accuracy) for single figurative form modelling and cross-figurative modelling in English and Spanish.



(a) Results between templates A and B.

(b) Results between templates A and C.

(c) Results between cross-lingual and in-lingual templates (A VS D).

Figure 5: Relative performance differences between different prompt templates.

pared to single figurative form models, we see that multitask modelling yields further improvements in hyperbole and idiom but hurts metaphor. Similarly, when combining information on both idioms and metaphors for ES, extra information about idioms hurts metaphor detection slightly, while extra information about metaphors helps idioms. We suggest two main reasons for these observations: (i) performance improvements in hyperbole and idioms are enhanced by the transfer of knowledge from metaphors; (ii) The low-quality idiom training data, as discussed earlier in this section, negatively impacts the accuracy of metaphor detection. While incorporating information from hyperbole data could potentially be beneficial, the limited amount of such data might not be enough to bring any benefit.

Cross-Lingual Knowledge Transfer We use the model trained in one language to run the zero-shot experiments on the other languages and model all languages jointly for each figure of speech. Figure 4 shows the results for cross-lingual experiments. Zero-shot has moderate detection accuracy on hyperbole and metaphor with scores greater than 68% for all languages, confirming that figurative knowledge is transferable across languages.

In idiom detection, it is unsurprising to see that zero-shot performs poorly, e.g. the accuracy of the DE model on EN idiom detection is only 29.5% considering that cultural specificities of idioms might hamper cross-lingual generalization more than for other figures of speech. Still, multilingual modelling brings performance improvements on most tasks in different languages, including idioms. Overall, figurative language detection can benefit from multilingual modelling, and the zero-shot technique can be used for hyperbole and metaphor detection when lacking resources in the target language, but not, in most cases, for idiom detection.

Impact of Prompt Although prompt learning has been shown an effective method for many NLP

Task Lang	Prompt Lang	#	Prompt Template
Italian	English	A	Which figure of speech does this text contain? (A) Literal. (B) [TASK]. | Text: [IT-text]
		B	Is there a(n) [TASK] in this text? | Text: [IT-text]
		C	Does this text contain a(n) [TASK]? | Text: [IT-text]
	Italian	D	Quale figura retorica contiene questo testo? (A) Letterale. (B) [TASK]. | Testo: [IT-text]

Table 7: Examples of different prompt templates. TASK and IT-text represent the placeholders for figure of speech (e.g. idiom) and the text in Italian, respectively.

tasks, it usually requires extensive prompt engineering on template design as it is sensitive to different tasks. Following Fu et al. (2022), we assess the prompt effect with different templates and languages. In Table 7, we show a set of cross-lingual and in-lingual prompt templates. In the cross-lingual prompt setting, all templates are written in English, here we experiment with two other templates (B and C) besides the one used in our main results (A). In the in-lingual prompt setting, instead, the language of the template is consistent with the task language. Template D in Table 7, for example, is an in-lingual prompt translated from A and used for Italian tasks.

Figure 5 shows the relative performance differences, where we subtract the performance of the model with other prompts from the model with prompt A. In the cross-lingual setting, we see that the performances of models with different prompt templates are very close, with an accuracy difference of less than 5 percentage points on all tasks except for EN and DE idiom using template A and C. Interestingly, the English template does not hurt performances in other languages (Figure 5(c)). These results suggest that a model based on prompt learning for multilingual multi-figurative language detection is not particularly sensitive to different templates.

6 Conclusions

We introduced a multilingual multi-figurative language understanding benchmark that focuses on sentence-level figurative language detection, involving three common figures of speech and seven languages. Based on prompt learning, we proposed a framework to unify the interrelated detection tasks across multiple figures of speech and languages using a PLM, while having no task- or language-specific modules. We further analyzed the generalization of the model across different figures of speech and languages.

Our unified model benefits from cross-lingual and cross-figurative knowledge transfer in sentence-level detection. It is natural to explore fine-grained detection at the word-level in future work, as well as language generation in multilingual and multi-figurative scenarios. This approach can also serve as a blueprint for the joint modelling of other interrelated tasks.

7 Limitations and Impact

While introducing a framework which deals with multiple languages and multiple figures of speech, this work is still only dealing with three figures of speech and seven languages. Many more phenomena and languages can still bring substantial challenges and insights if considered (once the data availability bottleneck is addressed). Also, we deal with figurative language as labelled at the sentence level, but the word level is also not only interesting but important for broader natural language understanding and could yield different insights than those observed in the present work.

We only mention in passing the influence that different cultural contexts have on figurative usages, and we make some observations on idioms, but this aspect would require a much bigger unpacking. We actually believe that (failure) of cross-lingual computational models can be an excellent diagnostic tool towards a finer-grained analysis of the interplay between culture(s) and figurative language.

We propose a successful method based on prompt learning and present experiments using a specific pre-trained model. Choosing different (and possibly larger) models and investigating even more than what we already do in this paper the influence of specific prompts would also be necessary to further generalise the efficacy of our approach.

Finally, as with most language technology, the limitations of our approach, also in terms of accuracy (especially for some phenomena and some languages), could lead to substantial inaccuracies which could be propagated in further processing. Considering that figures of speech are associated with emotional language, a word of warning is necessary regarding the direct deployment of our

models. We do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful consquences. We are open to any discussion and suggestions to minimise such risks.

Acknowledgments

This work was partly funded by the China Scholarship Council (CSC). The anonymous reviewers of ACL 2023 provided us with useful comments which contributed to improving this paper and its presentation, so we're grateful to them. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors. 2018. *Proceedings of the Workshop on Figurative Language Processing*. Association for Computational Linguistics, New Orleans, Louisiana.

George Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. volume 7812, pages 102–110.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

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, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask promptraining. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, and Tuhin Chakrabarty, editors. 2022. *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Sarah Harmon. 2015. Figure8: A novel system for generating and evaluating figurative language. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 71–77.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection

in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. Association for Computational Linguistics, Online.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying exaggerated language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034, Online. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.

Huiyuan Lai and Malvina Nissim. 2022. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5939–5954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* Just Accepted.

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, Florence, Italy. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 56–63, Sapporo, Japan. Association for Computational Linguistics.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Omid Rohanian, Shiva Taslimipoor, Richard Evans, and Ruslan Mitkov. 2018. WLV at SemEval-2018 task 3: Dissecting tweets in search of irony. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 553–559, New Orleans, Louisiana. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

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, Online. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against Neural Fake News*. Curran Associates Inc., Red Hook, NY, USA.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and Sec. 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3.2

☑ B1. Did you cite the creators of artifacts you used?
3.2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we used are freely available and were created for the same tasks that we use them for.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The datasets we used are freely available and were created for the same tasks that we use them for.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets are commonly used for this task.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.2, Table 2

C ☑ Did you run computational experiments?

4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
partly in Sec. 5.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
partly in Sec. 5.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.