Nut-cracking Sledgehammers: Prioritizing Target Language Data over Bigger Language Models for Cross-Lingual Metaphor Detection

Jakob Schuster
Linguistic Data Science Lab
Ruhr University Bochum
Germany
jakob.schuster@ruhr-uni-bochum.de

Katja Markert
Institute of Computational Linguistics
Heidelberg University
Germany
markert@cl.uni-heidelberg.de

Abstract

In this work, we investigate cross-lingual methods for metaphor detection of adjective-noun phrases in three languages (English, German and Polish). We explore the potential of minimalistic neural networks supported by static embeddings as a light-weight alternative for large transformer-based language models. We measure performance in zero-shot experiments *without access to annotated target language data* and aim to find low-resource improvements for them by mainly focusing on a k-shot paradigm. Even by incorporating a small number of phrases from the target language, the gap in accuracy between our small networks and large transformer architectures can be bridged. Lastly, we suggest that the k-shot paradigm can even be applied to models using machine translation of training data.

1 Introduction

Metaphors are a phenomenon of figurative language where meaning about a more abstract concept is expressed by applying it to a more concrete domain. According to cognitive linguistic theories by Lakoff and Johnson (1980), they are systematic linguistic instantiations of so-called Conceptual Metaphors. An example of a conceptual metaphor is `EMOTION IS LIQUID`, which manifests in expressions such as *bubbly personality, his anger boiled over* and *overflowing joy*. Other definitions describe metaphors as novel usages of words, in which the semantic preference of the syntactic arguments is violated. As an example, *to eat* prefers animate subjects and edible objects. The metaphor *The job ate his confidence* violates this preference (Wilks, 1975). Previous studies show metaphors to make up a substantial portion of natural language[1] and heavily influence decision-making in discourse (Thibodeau and Boroditsky, 2011), making their detection a valuable topic in NLP. Since conceptual metaphors are based around semantic concepts and not words, they are shared throughout similar cultures and can sometimes be directly translated (*Er griff mein Argument an - He attacked my argument*). In other cases, the same conceptual metaphor might exist in two languages, but is lexicalized differently. While a direct translation of *Seine Stimmung war im Keller - His mood was in the basement* could most likely still be understood, a more conventional phrasing would be *His mood plummeted* or *He was feeling down*. Therefore, metaphor detection across different languages is an interesting topic worth exploring. However, most annotated metaphor resources center on English.

In this paper, we investigate the application of modern zero-shot methods without access to annotated target language data for cross-lingual metaphor detection of adjective-noun phrases in three different languages. We go on to soften the zero-shot limitation and measure how smaller feed-forward models can become competitive to transformer-based systems, by incorporating a small number of target language phrases into the training process. Lastly, we apply the same few-shot paradigm to improve models which use machine-translated data and discuss the results.

2 Related Work

Previous works about metaphor detection were mostly monolingual and supervised. They often leveraged additional resources, static word embeddings and in more recent experiments pre-trained transformer models (Wilks et al. (2013), Do Dinh and Gurevych (2016), Choi et al. (2021)). The latter is currently the most commonly used option. In a shared task about metaphor detection in 2020 (Leong et al., 2020), more than half of all participants used some variation of a transformer architecture. Recent concerns regarding the alternative

[1] The VUA Metaphor Corpus by Steen et al. (2010) annotates around 12% as metaphoric.

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 98–106
September 11–12, 2023.

usage of static word embeddings for metaphor detection were also voiced for theory based reasons (Maudslay and Teufel, 2022). Work addressing cross-lingual metaphor detection includes Tsvetkov et al. (2014), where semantic features and word vectors were used to transfer English metaphor knowledge about adjective-noun or verb-subject-object phrases into Spanish, Farsi and Russian, Schneider et al. (2022), an unsupervised approach for a transfer from German to middle high German based on self-trained fasttext embeddings (Grave et al., 2018) and Sanchez-Bayona and Agerri (2022) who present first zero-shot results between their newly created Spanish corpus and the English VUAMC by Steen et al. (2010) by using XLM-RoBERTa (Conneau et al., 2019), a transformer based multilingual language model.

This leaves many possible approaches to cross-lingual transfer-learning for metaphor detection unexplored. A common way to allow for this transfer in other tasks is machine translation to convert training data from the source language into the target language or vice versa (Eger et al., 2018). Joulin et al. (2018a) provide a lightweight alternative to bigger transformer models, by aligning static fasttext embeddings across 44 languages through a cross-domain similarity local scaling criterion. While multilingual models do work in zero-shot scenarios, Lauscher et al. (2020) show the benefit of shifting to a k-shot scenario, in which small target language datasets of size k are incorporated into training. Similarly, Keung et al. (2020a) present findings which support that using a development set in the target language can improve performance by preventing catastrophic forgetting of multilingual knowledge during training. More recently, Large Language Models such as ChatGPT[2], are used as zero-shot or few-shot in-context learning systems (Laskar et al. (2023), Yuan et al. (2023)). ChatGPT is a model of the GPT-3.5 or GPT-4 series, which is trained through a reinforcement learning from human feedback component (Christiano et al., 2017) and also possesses multilingual knowledge.

3 Task and Data

This paper focuses on binary classification of the metaphoricity of adjective-noun tuples, since this setup had the most available data in several languages. In these phrases, the metaphoric meaning can stem from the conceptual transfer of either the meaning of the adjective (*stale idea*) or the noun (*economic slump*). We collected corpora of labeled phrases big enough for both training and testing in English, German and Polish. A small sample can be seen in Table 1.

Example Inputs	Gold Label
wet towel, old man, ...	0
broken home, cultural barrier, ...	1

Table 1: Example classification schema for the metaphor detection task of adjective-noun phrases. 1 indicates a metaphorical and 0 a literal meaning.

	Size	%M	#adj	ppa
DE	1677	25.5	297	5.6
EN	1968	50.0	668	2.9
PL	2052	50.4	241	8.5

Table 2: Comparison of the annotated source datasets. By measuring the simple attributes share of metaphoric phrases (**%M**), number of adjective types (**#adj**) and phrases per adjective (**ppa**), we can show how the different strategies result in different distributions.

The English corpus (Tsvetkov et al., 2014) is balanced for both classes, and consists of metaphor annotations of the 1000 most common adjectives and their co-occuring nouns in the TenTen Web corpus.[3] It has been filtered to exclude phrases which without context can be interpreted literally and metaphorically (e.g. *drowning students*).

The German corpus (Sick, 2020) follows the same annotation procedure as Tsvetkov et al. and is extracted from the German deTenTen13[4] corpus. The resulting dataset is not balanced between classes, but rather reflects the distribution of metaphoric tokens in natural language. The Fleiss' κ (Fleiss, 1971) measuring inter-annotator agreement is 0.34. Since this is a low IIA, we filter the corpus and only include phrases for which at least 4 of the 5 annotators agreed.

The Polish corpus (Mykowiecka et al., 2018) is constructed by preparing a list of metaphorical phrases and enriching it with additional common phrases in the National Corpus of Polish (Przepiórkowski and Patejuk, 2014), using the same adjectives. After we removed phrases that were labeled as *Both* metaphorical and literal,

[2] https://chat.openai.com/

[3] https://www.sketchengine.eu/ententen-english-corpus/

[4] https://www.sketchengine.eu/detenten-german-corpus/

the corpus is almost perfectly balanced (1018 metaphors and 1034 literal phrases).

Due to the similar collection strategies, we can observe examples of the same conceptual metaphors being present in every corpus. (EMOTIONALLY INDIFFERENT IS COLD: *cold justice*, *kalte Grausamkeit* and *zimna kalkulacja*). A comparison of all sources can be seen in Table 2. To even out the differences in size, we trim every corpus down to the size of DE, while keeping the overhang in a separate set for later experiments. We then perform a 70:15:15 train, dev, test split, resulting in 1173 phrases for training and 252 each for developing and testing. Since we use our own test splits, we have no previous results from literature to compare against.

4 Experiments

In this section, we describe a series of binary classification experiments of our collected phrases. Each experiment described is conducted for all six possible combinations of training and test splits of our three available languages. We prioritize stability of our results over ideal hyperparameters and aim to ensure a fair comparison. Therefore, in all following experiments, we incorporate early stopping, learning rate warm-up and report the average result of ten majority vote ensembles with seven seeds each.

4.1 Upper Bound

Previous work for similar semantic tasks have shown big gaps in performance between cross-lingual and monolingual set-ups (Nozza, 2021; Hsu et al., 2019). As an approximation of an achievable upper-bound for our cross-lingual models, we first conduct monolingual experiments with language-dependent BERT variations[5] and light-weight, fully connected feed forward neural networks, using fasttext word embeddings[6].

4.2 Zero-shot Models

The cross-lingual zero-shot experiments of this section are defined by the absence of annotated target language examples in the training dataset. We compare models of three different categories for cross-lingual zero-shot metaphor detection. The first category consists of networks powered by aligned **fasttext** word embeddings by Joulin et al. (2018b). We train three additional variations of this architecture:

- **fasttext+TrTr** and **fasttext+TrTe**, with translations of the training data into the target language or the test data into the source language.[7]
- **fasttext+TarDev** which employs a development set in the target language as proposed by Keung et al. (2020b). Using a development set in the target language can enable a checkpoint selection that best suits the test data.

The second category encompasses the two multilingual pre-trained transformer models **MBERT** (Devlin et al., 2018) and **XLM-R**[8], which are fine-tuned on the source language for the classification.

The final category describes a set of experiments, utilizing **ChatGPT** [9] as a classifier via prompting:

- **ChatGPT** is not given any additional information.
- **ChatGPT+ex** is provided with 20 random examples from the source language's training split before (In-Context Few-shot Learning)
- **ChatGPT+MIP** is provided with the (translated) Metaphor Identification Procedure by Group (2007) and asked for corresponding annotations (In-Context Instruction Learning)

Example prompts for all three ChatGPT methods can be found in the Appendix.

4.3 k-shot for Fasttext Models

Just as proposed by Lauscher et al. (2020), in this series we relax the zero-shot limitation to explore an inexpensive approach of mitigating the gap between cross-lingual and monolingual performance. We incorporate k randomly sampled data points from the target language's training, development or overhang data into the training process of the fasttext baseline. This sample is different

[5]All BERT variations are finetuned with a learning rate of 2e-5 and Adam's epsilon of 1e-8 for 8 epochs

[6]All our fasttext networks consist of three hidden layers ($h_1 = 300$, $h_2 = 150$, $h_3 = 50$), with a dropout chance of 5% and are trained for 5 epochs.

[7]We use the neural machine translation *Amazon Translate*, provided by the *Amazon Web Service*. It has to be noted that using a big NMT service such as Amazon Translate adds a hidden compute to all related experiments.

[8]XLM-R is finetuned with a learning rate and Adam's epsilon of 10e-5 for 6 epochs

[9]GPT-3.5 model of the May 3, 16k version with a temperature of 0.05

	huggingface transformer model	BERT ACC	fasttext ACC
DE	german-bert	**80.5**	79.3
EN	bert-base	**88.9**	83.7
PL	dkleczek/bert-base-polish-uncased-v1	85.9	**86.9**

Table 3: For each language, the used monolingual BERT model from the huggingface model hub and the accuracies produced by said monolingual BERT's experiments and the monolingual experiments using fasttext word embeddings.

for each seven seed ensembles and each set of size k is a subset of another set of a larger k. We report results for $k \in \{0, 10, 25, 50, 100, 200, 1173\}$, where $k = 1173$ shows the maximum achievable effect by including the whole training split of the target language.

4.4 k-shot for Translated Train Models

Lastly, we examine if even transformer-based language models, which are supported by machine translation, can still benefit from the k-shot paradigm. For this, we finetune monolingual BERT models in the target language on the translated training data as our $k = 0$ baselines. We then add additional k phrases of the training, development and overhang split of the target language. These are untranslated and authentical data points. To contrast their effect to the addition of more translated training data, we use the test split of the source language dataset to also finetune models with additional k translated phrases.

5 Results

In this section, we present the results of the experiments we conducted. For all of them, we report the accuracy on the test sets.

5.1 Upper Bound

Table 3 contains the results of the monolingual experiments. Overall, the BERT baselines outperform the fasttext model and the German dataset yields the lowest accuracy. However, this mainly serves as a potential upper-bound for the upcoming cross-lingual experiments.

5.2 Zero-Shot Models

Table 4 displays the accuracy of all zero-shot models. Generally, accuracy of all zero-shot systems varies across language pairs and models, with the inclusion of the German dataset seemingly often leading to worse results. Across all systems and languages, **ChatGPT+MIP** performed the best and achieves an average accuracy of 67%. On average, the other transformer models were able to outperform the plain fasttext architecture, albeit not for every language pair. When utilizing machine translation however, the models with translated training data nullified the gap to the transformer models in almost every pair, while the models with translated test data became worse overall. How dependent this behaviour is on the used translation service was not examined. We also observe that the inclusion of a development set in the target language does not bring a notable improvement to our fasttext architecture. This could be due to the small training data size, where not enough meaningfully different checkpoints are available for choosing. It is important to mention that all three of the categories feature models which performed worse than a random baseline. Models based on ChatGPT also display peculiar behaviour, with the additional information through examples of a source language seemingly weakening its predictive power. As expected, a comparison of Table 3 and Table 4 shows that transfer-learning across languages leads to a strong drop in performance for this task.

	DE ->EN	DE ->PL	EN ->DE	EN ->PL	PL ->DE	PL ->EN	avg.
fasttext	48.2*	60.0	63.6	68.3	59.0	62.9	60.3
fasttext+TrTr	59.9	**65.8**	47.6*	**72.2**	**63.1**	68.6	62.8
fasttext+TrTe	44.4*	55.1	56.7	65.4	60.7	68.6	58.4
fasttext+TarDev	48.1*	59.9	65.4	68.2	59.5	62.3	60.5
XLM-R	59.3	60.8	62.5	67.3	49.5*	74.6	62.3
MBERT	60.2	63.0	**66.3**	66.8	49.4*	68.0	62.2
ChatGPT	57.1	63.1	62.6	63.1	62.6	57.1	60.9
ChatGPT+ex	56.7	63.1	55.1	68.6	53.1	48.0*	57.4
ChatGPT+MIP	**77.3**	65.0	57.9	65.9	56.0	**74.6**	**66.1**

Table 4: Report of all the zero-shot baseline systems for every available language pair and the average across all language pairs. For **ChatGPT**, there is no actual source language from which we transfer knowledge to a target language. Therefore, the results for the two source languages are always identical. We mark every model worse than a random baseline with *.

5.3 k-shot for fasttext Models

Figure 1 displays heatmaps of the change in accuracy for all language pairs for rising k. Identically to the findings of Lauscher et al., we can observe a static incline of accuracy with rising k. Combinations that performed poorly in zero-shot rapidly improve, even for small values of k. On average, fasttext outperforms MBERT and XLM-R at $k = 25$ and even our best ChatGPT+MIT model for $k = 100$. Comparisons of $k = 200$ and $k = 1173$

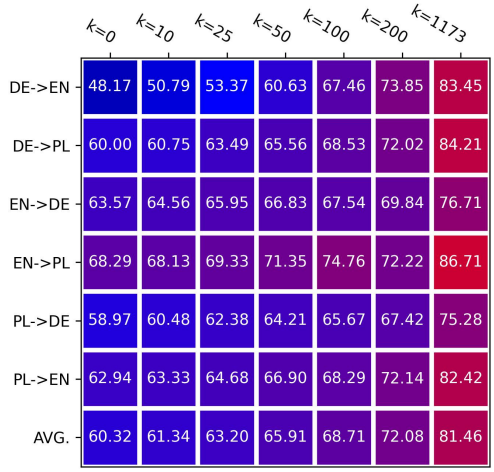

Figure 1: Heatmaps of accuracy across language pairs and the average across all pairs for rising values of k for vanilla fasttext models

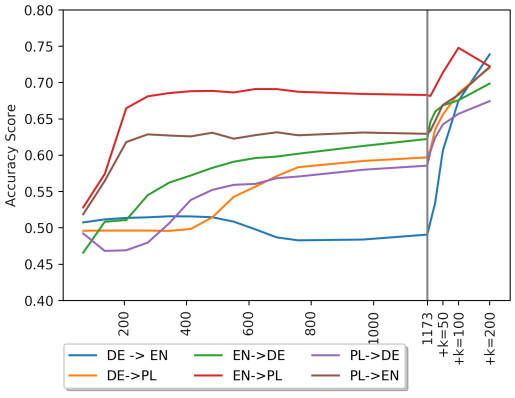

Figure 2: Learning curve of all language pairs for increasing source language training set size and additional k-shot data. The vertical line signifies the initiation of adding k-shot points to the complete training data set. Ten models area averaged for the reported accuracy score.

show that by only using 17% of the target language data, we can already obtain more than half of its potential increase in accuracy.

Another visual representation of the effectiveness of softening of the zero-shot limitation can be seen in Figure 2. The plot shows every available language pair's learning curve. By starting at an empty set and continuously adding data points of the source language to the training data set, it can be measured how much a model profits from more training data from that source. After adding the whole training data set, we then shift to further adding k-shot points. The vertical line indicates the point of this shift. It is evident that every model's learning curve slope gets steeper when switching

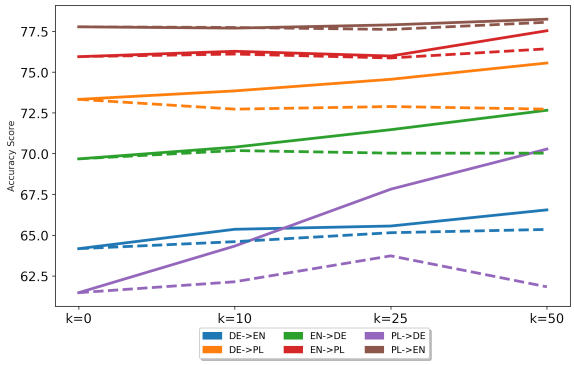

Figure 3: Comparison of accuracy across language pairs for rising values of k. Using the monolingual BERT models listed in Table 3 and additional translated training data (dashed line) or authentic data from the target corpus (solid line).

to the k-shot paradigm. This applies to pairs where the accuracy appears to plateau (EN $\rightarrow$ PL, PL $\rightarrow$ EN), to DE $\rightarrow$ EN, which seems to not improve at all and is outperformed by a random baseline, and to models which were past their stronger initial incline, but were still slightly improving.

5.4 k-shot for Translated Train Models

The zero-shot models of this experiment, while being not as light-weight due to the compute of BERT and the NMT, are roughly comparable to fasttext's $k = 100$ and $k = 200$ models in accuracy. The overall best models presented in this paper were obtained by this method for $k = 50$ for authentic k-shot, reaching an average accuracy of 73.47%. When comparing both methods in Figure 3, we can note that the models do not noticeably improve by additional translated training data. The same does not apply to the k-shot set of authentic data, where we observe a similar improvement to the k-shot experiments with fasttext - stronger improvements for models with worse performance in zero-shot.

6 Discussion

Impact of Machine Translation By looking at our translated data, we try to explain why the translation based zero-shot BERT experiments benefited more from the translations than the fasttext baselines. By our choice of method, we end up with translations of individual data points where the two word adjective-noun pair structure is lost (*warmer Milchschokoladenton* to *warm milk chocolate tone*, *crushed stone* to *Schotter*). By automatically POS-tagging the translated test data with spaCy (Honnibal and Montani (2017)), we measure these devia-

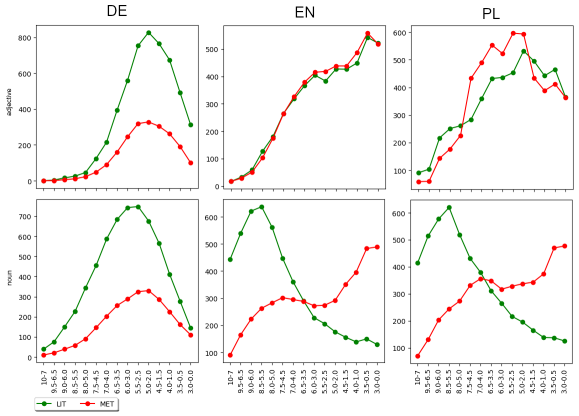

Figure 4: Distribution of abstractness/concreteness from the ratings provided by (Köper and Schulte im Walde, 2017) for the different corpora, separated into adjectives (top row) and nouns (bottom row). Words are rated on a scale of 0 to 10. Lower scores are given to abstract words (*irresponsibly*), higher to concrete words (*razor blade*).

tions in syntactic structure. Depending on language pair and direction, they make up between 7.5% and 38.4% of the translations. In general, length difference of the translations is less of a problem for the scalable transformer models, than for our 600 dimensional, fixed length neural network. This can be circumvented by using recurrent neural networks. Analogue to our findings of preferring simple models to large ones for this task, the same could possibly apply to the translation methods, since statistical or dictionary based methods could lead to less deviation in syntax and therefore to better results. This deviation can also explain how our models profit more from the authentic k-shot data, since it better represents the test data. We also observed the NMT already implicitly performing metaphor detection to better lexicalize concepts in the target language (*schwieriger Spagat* to *difficult balancing act* instead of a less conventional *difficult split* for the conceptual mapping `LIFE IS A SEQUENCE OF MOTION`).

Performance Difference Between Languages In order to try and explain the differences in performance for the individual language pairs, we investigated the semantic composition of the corpora. Using abstractness/concreteness ratings from Schulte im Walde (2022), we display the distribution of abstractness for the adjectives and nouns of our datasets in Figure 4. The DE corpus differs heavily, by having similar distributions of abstractness for metaphoric and literal words. In comparison, EN and PL contain more concrete nouns in literal and abstract nouns in metaphoric phrases. This is more in line with work by Turney et al. (2011), Tsvetkov et al. (2013) and Schulte im Walde (2022), where abstractness served as a classification feature and can serve as an indicator for the lower performance on the German test set.

Impact of k-shot Selection To gain insight into the effect of the selection of the k datapoints, we look at the performance of individual ensemble seeds with different k-shot sets. We investigate the intuitive connection between the seed's performance and the coverage of the test set adjectives by the k-shot data and show an exemplary scatter plot for our fasttext model and EN $\rightarrow$ PL in Figure 5. While larger values of k lead to a better performance and also naturally to a higher coverage of adjectives, when looking the distribution inside a cluster of k, there seems to be no strong connection. This makes the k-shot paradigm robust, since no knowledge about the word content of the test dataset is therefore needed. The plot also shows the k-shot data to improve both the detection of metaphors and the detection of literals. It is worth noting that the variance in performance appears to be higher for smaller values of k, with some poor performing outliers, while higher values of k produce more stable results.

Multiple efforts have been made to enhance the selection of k-shot data, similar to Lauschers selection based on length. Experiments based on attributes such as class label, frequency, distance of the data points in the vector space or other small handcrafted feature vectors were all unreliable and too dependent on the language pair and k. However, based on the variance in performance for smaller k, we can not rule out the potential benefit of a more sophisticated selection process and leave it for future work.

7 Conclusion and Future Work

The findings of this paper serve to reinforce the idea that larger language models are not always inherently superior at every task and should therefore not automatically be considered the default choice. We have shown how primitive fasttext models can be competitive with large transformer based language models for syntactically trivial but semantically complex tasks such as cross-lingual metaphor detection of adjective-noun phrases. Furthermore, these small models can easily be enhanced to out-

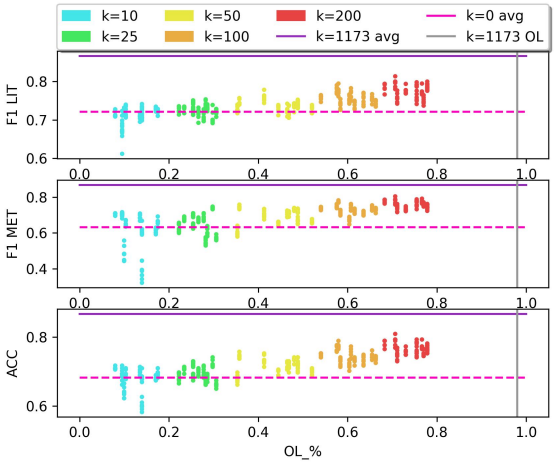

Figure 5: Scatter plot of the different fasttext seeds for rising k and exemplary for EN to PL. We distinguish between the F1 score of literals, metaphors and accuracy. The points are coloured for k and scattered by their percentage of seen adjectives of the test data through the k-shot data (OL%). Line plots are provided for the average performance and adjective overlap when including the whole training data and the average performance for zero-shot.

perform their substantially larger competitors by softening the zero-shot limitation and including small amounts of data from the target language. Based on our experiments, we recommend using k-shot data as a quick and cost-effective measure, over upscaling to a substantially bigger language model. This has the benefit of a computationally less demanding training environment, almost immediate results and a more environmentally friendly model.[10] The claim about static word embeddings being ill-suited for metaphor detection (Maudslay and Teufel (2022)) can not be validated by our findings, since they performed similarly to the contextualized representations. Additionally, we showed that the k-shot paradigm can also benefit architectures utilizing automatically translated training data.

Investigation of the peculiar behaviour of ChatGPT's dependence on the prompts, as seen in Table 4, would be interesting but was beyond the scope of this paper. Finally, we leave expanding the k-shot experiments to sequence labeling for metaphor detection open for future work.

[10]Finetuning the BERT models took more than 3 hours, while our training of our fasttext models concluded after just 30 seconds.

Acknowledgements

We thank Theresa Sick for providing us with the corpus of annotated adjective-noun phrases in German and Pola Stawiany for providing the Polish translation of the Metaphor Identification Procedure and insight into the Polish corpus and translated phrases.

References

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! *arXiv preprint arXiv:1807.08998*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-Yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018a. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018b. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020a. Don't use english dev: On the zero-shot cross-lingual evaluation of contextual embeddings. *arXiv preprint arXiv:2004.15001*.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020b. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Chee Wee (Ben) 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, Online. Association for Computational Linguistics.

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

Agnieszka Mykowiecka, Malgorzata Marciniak, and Aleksander Wawer. 2018. Literal, metphorical or both? detecting metaphoricity in isolated adjective-noun phrases. In *Proceedings of the Workshop on Figurative Language Processing*, pages 27–33.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Adam Przepiórkowski and Agnieszka Patejuk. 2014. Koordynacja leksykalno-semantyczna w systemie współczesnej polszczyzny (na materiale narodowego korpusu języka polskiego). *Język Polski*, pages 104–115.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2022. Metaphor detection for low resource languages: From zero-shot to few-shot learning in Middle High German. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 75–80, Marseille, France. European Language Resources Association.

Sabine Schulte im Walde. 2022. Figurative language in noun compound models across target properties, domains and time. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, page 1, Marseille, France. European Language Resources Association.

Theresa Sick. 2020. Metaphor detection in german adjective-noun-pairs. B.A. thesis, Heidelberg University.

Gerard Steen, Lettie Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. A method for linguistic metaphor identification from mip to mipvu preface. *Method For Linguistic Metaphor Identification: From MIP To MIPVU*, 14:IX–+.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

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.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44, Atlanta, Georgia. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Appendix

A ChatGPT Prompts

role	content
system	You are a multilingual metaphor detection system. You classify incoming adjective-noun phrases acording to their metaphoricity. Returning 1 if a phrase is a metaphor and returning 0 if a phrase is literal. You return your answers in JSON format, with the prediction at attribute 'label'.
user	[{"id": 1,"phrase": "grünes Landschaftsparadies"}, {"id": 2,"phrase": "knapper Fragenbogen"}, {"id": 3,"phrase": "jährlicher Fleischverzehr"}, {"id": 4,"phrase": "feine Grilladen"}, {"id": 5,"phrase": "amerikanische Erforscher"}...]

Table 5: Example prompt for the **ChatGPT** model

role	content
system	You are a multilingual metaphor detection system. You classify incoming adjective-noun phrases acording to their metaphoricity. Returning 1 if a phrase is a metaphor and returning 0 if a phrase is literal. You return your answers in JSON format, with the prediction at attribute 'label'.
user	[{"id": 1,"phrase": "unreadable face"}, {"id": 2,"phrase": "drowsy heat"}, {"id": 3,"phrase": "turbulent water"}, {"id": 4,"phrase": "smokey eyes"}, {"id": 5,"phrase": "metallic surface"}...]
assistant	[{"id": 1,"phrase": "unreadable face", "label": 1 }, {"id": 2,"phrase": "drowsy heat", "label": 1 }, {"id": 3,"phrase": "turbulent water", "label": 0 }, {"id": 4,"phrase": "smokey eyes", "label": 1}, {"id": 5,"phrase": "metallic surface", "label": 1 }...]
user	[{"id": 1,"phrase": "grünes Landschaftsparadies"}, {"id": 2,"phrase": "knapper Fragenbogen"}, {"id": 3,"phrase": "jährlicher Fleischverzehr"}, {"id": 4,"phrase": "feine Grilladen"}, {"id": 5,"phrase": "amerikanische Erforscher"}...]

Table 6: Example prompt for the **ChatGPT+ex** model

role	content
system	You are a multilingual metaphor detection system. You classify incoming adjective-noun phrases according to their metaphoricity based on the Metaphor Identification Procedure. Returning 1 if a phrase is a metaphor and returning 0 if a phrase is literal. You return your answers in JSON format, with the prediction at attribute 'label'. This is the Metaphor Identification Procedure: 1. Read the text to get a general understanding of the meaning 2. Determine the lexical units 3a. Establish the contextual meaning of the unit 3b. Determine if it has a more basic meaning. Basic meaning 'more concrete, body-related, more precise, historically older; not necessarily the most frequent meaning! Does the contextual meaning contrast with the basic meaning but can it be understood in comparison with it? 4. If yes, mark the unit as metaphorical.
user	[{"id": 1,"phrase": "grünes Landschaftsparadies"}, {"id": 2,"phrase": "knapper Fragenbogen"}, {"id": 3,"phrase": "jährlicher Fleischverzehr"}, {"id": 4,"phrase": "feine Grilladen"}, {"id": 5,"phrase": "amerikanische Erforscher"} ...]

Table 7: Example prompt for the **ChatGPT+MIP** model