Cross-Lingual Metaphor Detection for Low- to High-Resource Languages

Anna Hülsing
University of Hildesheim
anna.huelsing@uni-hildesheim.de

Sabine Schulte im Walde
University of Stuttgart
schulte@ims.uni-stuttgart.de

Abstract

Research on metaphor detection (MD) in a multilingual setup has recently gained momentum. As for many tasks, it is however unclear how the amount of data used to pretrain large language models affects the performance, and whether non-neural models might provide a reasonable alternative, especially for MD in low-resource languages. This paper compares neural and non-neural cross-lingual models for English as the source language and Russian, German and Latin as target languages. In a series of experiments we show that the neural cross-lingual adapter architecture MAD-X performs best across target languages. Zero-shot classification with mBERT achieves decent results above the majority baseline, while few-shot classification with mBERT heavily depends on shot-selection, which is inconvenient in a cross-lingual setup where no validation data for the target language exists. The non-neural model, a random forest classifier with conceptual features, is outperformed by the neural models. Overall, we recommend MAD-X for metaphor detection not only in high-resource but also in low-resource scenarios regarding the amounts of pretraining data for mBERT.

1 Introduction

Song titles such as *Life is a Highway* are prominent examples of how we use metaphors in our everyday life. But songs are by far not their only habitats: on average and across domains, metaphors can be found in every third sentence (Shutova and Teufel, 2010). Lakoff and Johnson (1980) define a conceptual metaphor as "understanding one conceptual domain [A] in terms of another conceptual domain [B]" (Kövecses, 2010). In the above example, the domain *Life* (A) is understood in terms of the domain *Journey* (B). Detecting whether or not a word or expression is a metaphorical linguistic expression (i.e. whether or not it is used metaphorically) is vital for many NLP applications, such as sentiment analysis, machine translation, information extraction, and dialog systems, cf. Tsvetkov et al. (2014). Metaphor detection (MD) can further support automatic essay scoring (Beigman Klebanov et al., 2018), schizophrenia detection (Gutiérrez et al., 2017), and propaganda identification (Baleato Rodríguez et al., 2023).

Many efforts have been made to tackle the task of metaphor detection (MD),[1] and successfully so: close-to-human performance was reached by systems using large pretrained language models like BERT (Devlin et al., 2019) for English datasets containing single sentences with a metaphorical expression (Ma et al., 2021). For a long time, Tsvetkov et al. (2014) were the only ones to perform MD cross-lingually, namely for Spanish, Russian and Farsi. Only recently, Aghazadeh et al. (2022) and Lai et al. (2023) addressed metaphor detection in a multilingual setup with the same languages as Tsvetkov et al. (2014). Whereas Aghazadeh et al. (2022) focused on probing metaphoricity within the transformer layers, Lai et al. (2023) used a template-based prompt learning approach to MD. These multilingual MD approaches focus on languages where large amounts of data are available for pretraining. Insights are missing, however, on whether or not large language models are also suitable for MD in languages with small amounts of pretraining data.

The current study addresses this bottleneck and compares neural and non-neural cross-lingual models for detecting metaphors in languages with varying degrees of pretraining data, including the low-resource language Latin.

Our metaphor detection focuses on word-based classification, as in the following example from the metaphor dataset by Tsvetkov et al. (2014):

(1) Actions talk even louder than phrases.

[1] See Shutova (2015), and Tong et al. (2021) for two prominent surveys.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 22–34

Language	# Wikipedia articles
English (source)	$\approx 6.7m$
German	$\approx 2.8m$
Russian	$\approx 1.9m$
Latin	$\approx 0.1m$

Table 1: Amount of articles in millions (m) regarding the four languages used in the current study. The numbers are taken from https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed 25 Sep. 2023. Altogether, mBERT was pretrained on Wikipedia articles from 104 languages.

We define a binary classification task to detect whether or not the underlined target word is used metaphorically in the given context. For zero- and few-shot classification we apply multilingual BERT (mBERT) (Devlin et al., 2019) and the adaptation method MAD-X (Pfeiffer et al., 2020b), which have shown state-of-the-art results for e.g. named entity recognition and question answering. As our non-neural model, we apply a random forest classifier (Breiman, 2001), as random forest classifiers generally perform well in low-resource scenarios (Tsvetkov et al., 2014). Our model utilizes a vector space model and conceptual features (abstractness and supersenses) – similarly to the model introduced by Tsvetkov et al. (2014).

As for target languages, we investigate modelling performances for German, Russian and Latin, because the amount of data used to pretrain mBERT varies greatly across these three languages (see Table 1). Whereas German and Russian are not considered low-resource languages in terms of pretraining data, we simulate low-resource conditions and explore the influence of different amounts of pretraining data by using as little as 20 instances or no labelled data at all from the target languages for training, and no data at all for validation. Latin, on the other hand, is a low-resource language in terms of pretraining data and in terms of labelled training data. English as a high-resource language was used as the source language for cross-lingual transfer.

Contributions. The main contribution of this paper is a comparison and a series of insights regarding cross-lingual neural and non-neural models for MD in languages with high-to-low degrees of pretraining data, i.e., German, Russian and Latin. More specifically, **1)** we find that with default hyperparameters, zero-shot mBERT performs best: results are above a majority vote baseline for all three target languages. **2)** MAD-X performs best when hyperparameter-tuning is carried out or large amounts of source language training data are used. **3)** We show that few-shot mBERT depends largely on shot-selection, which cannot be carried out in a low-resource environment where no validation data exists. **4)** Overall, the non-neural model is outperformed by the neural classifiers, and we recommend using MAD-X with suitable hyperparameters for MD in languages with both large and little amounts of data used for pretraining mBERT.[2]

2 Related Work

Metaphor Detection. Turney et al. (2011) were among the first to apply insights from cognitive linguistics to their MD model, i.e., exploiting that metaphors transfer knowledge from a concrete domain to an abstract domain (Lakoff and Johnson, 1980). Since metaphoricity is correlated with the degree of contextual abstractness, the authors used abstractness scores of context words as features in a logistic regression model.

The idea of "conceptual features" also inspired Tsvetkov et al. (2014), who used abstractness scores, imageability scores and semantic supersenses as classification features. Whereas Turney et al. (2011) focused on English data only, Tsvetkov et al. (2014) trained on English data and then evaluated the model cross-lingually on Spanish, Farsi and Russian. Their model represents the basis for the random forest classifier used in our experiments. Köper and Schulte im Walde (2016) focused on MD for German particle verbs. They also used 1) abstractness and imageability ratings as well as 2) scores indicating the distributional fit of particle verbs with regard to base verb contexts. In addition, they used 3) unigram context words and 4) noun clusters as features.

Do Dinh and Gurevych (2016) were the first to use a neural model architecture for MD, namely a multilayer perceptron with word embeddings. Their approach performed comparable to existing models without requiring feature engineering. Dankers et al. (2019) explored the relationship between metaphors and emotions by building several multi-task learning models. The best performing architecture made use of BERT embeddings used as input to a multilayer perceptron or to additional attention layers. They reached state-of-the-art results in 2019 for both metaphor and VAD prediction.

[2]The code can be accessed here: https://github.com/AnHu2410/MD_crosslingual.

Su et al. (2020) transformed word-based metaphor detection into a reading comprehension problem; their approach, DeepMet, was the most successful model in the 2020 metaphor detection shared task (Leong et al., 2020). Ma et al. (2021) fine-tuned BERT for MD. To perform word-based binary metaphor classification, they copied the input sentence and masked the target word. The original sentence and the masked copy were used as input for a sequence classification task. The BERT model then predicted whether the two sentences appeared in the same context; if yes, they predicted a literal usage of the masked word; otherwise they predicted a metaphorical usage. They also performed sentence-level classification and sequential labelling of metaphorical expressions. Their results showed an increase over previous state-of-the-art models. We use their word-based classification approach for the mBERT-based classifiers in our experiments. While their focus was on English, we use it in a multilingual setup.

Li et al. (2023) exploited the fact that many datasets are based on the Metaphor Identification Process (MIP; Pragglejaz Group, 2007), where a word is annotated as metaphorical if its contextual meaning is dissimilar to its "more basic meaning" (among further criteria). While prior models (such as MelBERT by Choi et al. 2021) grounded on MIP use decontextualized representations of the target word, Li et al. (2023) successfully gathered the representation of the target word from sentences where it was used literally.

Cross-Lingual Representations. Vulić and Moens (2013) proposed a bootstrapping method to create bilingual vector spaces from non-parallel data. Usually, a high-dimensional vector in a feature vector space uses context features as dimensions. For the proposed bilingual vector space, these features consisted of translation pairs. This method can be applied to any language pair.

Multilingual BERT (Devlin et al., 2019) was pretrained on data from 104 languages. Lauscher et al. (2020) pointed out limitations of large multilingual pretrained language models by demonstrating that these models do not transfer knowledge well for low-resource target languages (i.e. languages with small pretraining corpora) and for distant language pairs. They showed that first fine-tuning on large amounts of data and then continuing fine-tuning with very few examples from the target language considerably improves results across all languages and tasks. The current paper investigates whether these findings also apply to MD. Pfeiffer et al. (2020b) tried to mitigate problems of multilingual language models targeting low-resource languages by using an adaptation method, i.e. by inserting small amounts of trainable weights into an existing pretrained model (see Section 4). We also apply these Multiple ADapters for Cross-lingual transfer (MAD-X) to MD in our experiments.

3 Datasets and Preprocessing

Source Language. We used the dataset from Tsvetkov et al. (2014) as our basic English training dataset. It is based on the TenTen[3] Web Corpus, contains 222 instances, and is balanced. This basic training dataset was previously used by Tsvetkov et al. (2014) for evaluation. In the course of our experiments we augmented the amount of training data by adding the imbalanced dataset by Mohammad et al. (2016), which consists of 1,639 instances. The augmented version comprises 1,861 instances.[4]

Target Languages. Tsvetkov et al. (2014) also provide the Russian dataset that we used for evaluation, which is balanced, consists of 240 instances, and is based on the TenTen Web Corpus. For evaluation in German, we used the MD dataset provided by Köper and Schulte im Walde (2016), which is based on the web corpus DECOW14AX (Schäfer and Bildhauer, 2012) and where the target words are particle verbs. To balance the dataset, we reduced the original dataset from Köper and Schulte im Walde (2016) to 896 metaphorical and 896 literal instances.

For our Latin dataset we used the Lexham Figurative Language of the New Testament Dataset (Westbury et al., 2016), which is published in the Logos[5] Bible Software. It shows passages from the New Testament (we used the American Standard Version of the Bible), and highlights the metaphors in each verse. We extracted 100 sentences, of which 50 were annotated as metaphorical and 50 were annotated as literal. As the metaphors were annotated in the English Bible text, we then manually searched for the Latin translations in the Vulgate[6].

[3] https://www.sketchengine.eu/

[4] For the random forest classifier, only a subset of the dataset by Mohammad et al. (2016) was used for augmentation, because lemmatized subjects, verbs and objects had to be annotated, but this annotation was available only for 100 instances.

[5] https://www.logos.com

[6] https://vulgata.info/index.php?title=Kategorie:BIBLIA_SACRA.

The first author of this paper, a classical philologist, ensured that the metaphors found in the English texts correspondingly occurred in the Latin texts, i.e. that the American Standard Version did not introduce metaphors that were not present in the Vulgate.[7]
Below we provide two example sentences for each dataset, together with the respective categorization into metaphorical vs. literal.

- English (Tsvetkov et al., 2014, source):

  (2) The twentieth century saw intensive development of new technologies. → *metaphorical*

  (3) The young man shook his head. → *literal*

- English (Mohammad et al., 2016, source):

  (4) This young man knows how to climb the social ladder. → *metaphorical*

  (5) Did you ever climb up the hill behind your house? → *literal*

- Russian (Tsvetkov et al., 2014, target):

  (6) Бедность давит на людей.[8] (translation: "Poverty weighs on people.") → *metaphorical*

  (7) Повар варит суп на кухне.[9] (translation: "The cook cooks soup in the kitchen.") → *literal*

- German (Köper and Schulte im Walde, 2016, target):

  (8) Dort wird das Wasser aufgestaut und an Nimroz verkauft. (translation: "There, the water is dammed up and sold to Nimroz.") → *literal*

  (9) Über die Zeit hatte sich in ihnen Sehnsucht und Verlangen aufgestaut. (translation: "Over time, longing and desire had dammed up inside them.") → *metaphorical*

- Latin (Westbury et al., 2016, target):

  (10) Et venerunt, et impleverunt ambas naviculas, ita ut pene mergerentur. ("And they came, and filled both the boats, so that they began to sink.") → *literal*

  (11) Et dixerunt ei: Quia heri hora septima reliquit eum febris.("They said therefore unto him, Yesterday at the seventh hour the fever left him.") → *metaphorical*

We preprocessed all datasets such that the original sentence was available, as well as a copy of the original sentence, where we replaced the target word by the [MASK]-token. These two sentences were then further preprocessed by the HuggingFace[10] tokenizer pipeline. In addition, the random forest classifier required the target word (a verb) and its dependent subject and object as lemmas, which we annotated in cases where the information was missing. Figure 1 illustrates an example of input and output across models.

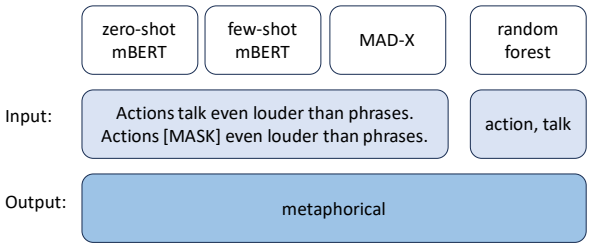

Figure 1: Example input and output of our models.

4 Models

For zero-shot and few-shot classification, we used mBERT (Devlin et al., 2019). For zero-shot classification, we fine-tuned the pretrained language model for MD on the source language data and used this model for predictions in all three target languages. For few-shot classification, we first fine-tuned mBERT on source-language training data, and then fine-tuned it again on a small amount of target language data (see Lauscher et al., 2020). Additionally, we applied MAD-X (Pfeiffer et al., 2020b), which consists of three types of adapters: language adapters, task adapters and invertible adapters. For this method, the pretrained model was frozen and two language adapters were trained on a masked language modelling task: one adapter was trained on unlabelled data from the source language, and one on unlabelled data from

[7] The German and Latin dataset are available here: https://github.com/AnHu2410/MD_crosslingual

[8] Transliteration: *Bednost' davit na lyudey.*

[9] Transliteration: *Povar varit sup na kukhne.*

[10] https://huggingface.co/docs/transformers/main_classes/tokenizer

the target language. Then, the source language adapter was inserted in addition to the task adapter, and the latter was trained on labelled data from the source language. Finally, inference was performed by plugging in the target language adapter and the (language-agnostic) task adapter. The invertible adapters were plugged in simultaneously with the language adapters, but come with a slightly different architecture because they adapt the embeddings, while the language and task adapters were inserted into each transformer layer. For all neural models we utilized the word-based MD method by Ma et al. (2021), where the original sentence and a copy of that sentence with the masked target word were used as input for sequence classification.

As our non-neural model we replicated the random forest classifier by Tsvetkov et al. (2014). This model contains three feature types: 1) abstractness and imageability scores, which Tsvetkov et al. (2014) generated on the basis of the MRC ratings by Wilson (1997), 2) supersenses, i.e., "coarse semantic categories", where a word can belong to several synsets in WordNet (Fellbaum, 1998), each of which is associated with several supersenses. We created a feature vector with these supersenses as dimensions, e.g., the noun "head" occurs in 33 synsets, 3 of which are related to the supersense *noun.body*. The dimension corresponding to the supersense *noun.body* then receives 3/33 (example taken from Tsvetkov et al., 2014). 3) Further features were produced with the vector space model by Faruqui and Dyer (2014). This model utilizes multilingual information in order to generate similar vectors for synonymous words. All these features were extracted from the target word – in our case, a verb – and from its dependent subject and object. For cross-lingual inference, the model relies on one-to-many translations: all translations were given for a target language word, and the scores obtained for the translations were averaged (see Tsvetkov et al., 2014). For translation, we used Word2Word by Choe et al. (2020).

5 Experiments and Results

5.1 Experimental Setup with Basic and Augmented Training Data

We used the basic English dataset by Tsvetkov et al. (2014) for training, and the target language datasets for Russian, German and Latin for evaluation (see Section 3). Then we explored how each of the following cross-lingual classifiers performed on each of the target languages: zero-shot mBERT (mB0); few-shot mBERT with a second fine-tuning on 20 instances of target language data (mB20)[11]; MAD-X; and the random forest classifier (RF).

As hyperparameters for zero- and few-shot mBERT in this basic experimental setup we used the default hyperparameters from Huggingface (Wolf et al., 2020), namely a batch-size of 8, a learning rate of 5e-5, and 3 training epochs. As hyperparameters for MAD-X we used those mentioned by Pfeiffer et al. (2020b): a learning rate of 1e-4, a batch-size of 8 and 100 training epochs. As hyperparameters for the random forest classifier we used those from scikit-learn 1.2 (Pedregosa et al., 2011), namely 100 estimators, no max-depth limit, and Gini as split criterion. We repeated the runs for three different seeds in order to simulate the variance of results achieved on different GPU machines, and report the mean F1-scores as well as the standard deviation (SD). We ran the experiments on an AMD EPYC 7282 16-Core Processor with 32 threads and NVIDIA RTX A6000 GPUs[12]. Our baseline predicts all instances to be metaphorical.

The results for the basic training dataset are presented in the left panel of Table 2. Zero-shot mBERT (mB0) outperformed the baseline for all three languages, while the results for the other three models were all similar or lower (with the exception of mB20 for Russian), and the results for Latin even dropped below the baseline. The random forest classifier produced results lower than mB0.

In order to investigate whether or not the small amount of training data (222 instances) could be responsible for the partly low results, we augmented the basic training data with data from Mohammad et al. (2016) to 1861 training instances, and repeated the experiments. The results are presented in the right panel of Table 2. For mB0, Russian showed slightly higher F1-scores, while the other two languages showed lower F1-scores compared to the basic training dataset. mB20 only achieved a performance comparable to the baseline (except for German). For the random forest classifier the results improved for Russian but remained the same for German and Latin. MAD-X clearly profited from the augmented training data.

[11] The 20 instances are taken from the test datasets for mB20, so here the test datasets are slightly reduced in comparison to the test datasets used for the other experiments.

[12] Training times were for the most part shorter than 10 minutes. The only exception was the training with augmented training dataset for MAD-X with 100 epochs (<30 minutes).

	basic training dataset			augmented training dataset		
	ru	**ge**	**la**	**ru**	**ge**	**la**
baseline	66.7	66.7	66.7	66.7	66.7	**66.7**
mB0	81.1 ±6.9	**77.1** ±1.6	**69.6** ±1.9	82.8 ±14.0	72.5 ±5.2	66.1 ±1.0
mB20	**82.0** ±2.3	67.3 ±1.2	62.1 ±0.0	66.9 ±37.3	70.9 ±3.9	62.2 ±0.8
MAD-X	68.3 ±10.5	64.2 ±10.7	42.0 ±21.3	**87.6** ±2.1	**75.2** ±0.3	63.3 ±3.6
RF	78.6 ±0.7	71.2 ±0.7	66.7 ±1.5	86.2 ±0.7	71.3 ±0.5	66.5 ±0.3

Table 2: Mean F1-scores for verbal MD across three runs with different seeds (± SD) for hyperparameters with the basic and the augmented training dataset and across our target languages Russian (ru), German (ge) and Latin (la).

5.2 Few-Shot Classifier: Shot-Selection

Even though Lauscher et al. (2020) showed that few-shot fine-tuning improves the performance of using zero-shot mBERT, the results obtained in our experiments did not improve with a second round of fine-tuning with 20 target language instances (except for Russian when using the basic training dataset). We therefore investigated shot-selection by selecting five different randomly selected shots instead of one randomly selected shot as in the previous experiments. The results for using default hyperparameters[13] and the basic training dataset are shown in Table 3. While the mean scores are lower than for the best-performing other models, the maximum scores were competitive; SD was rather high across all languages. We manually checked whether the successful shots exhibit specific features in comparison to the non-successful shots, but no pattern could be identified.

	max.	**mean**	**SD**
ru	87.3	76.3	15.1
ge	80.9	75.2	6.0
la	66.7	51.8	29.0

Table 3: Maximum and mean F1-scores as well as SD for using five different shots of the target language datasets for the second fine-tuning of mBERT (default hyperparameters, basic training dataset).

5.3 MAD-X: Hyperparameter-Tuning

As preliminary experiments have shown that MAD-X heavily relies on suitable hyperparameters, as a next step hyperparameter-tuning[14] was carried out. Given that in the cross-lingual setup no validation data for the target language exists, we explored whether using a dataset from the source language English for validation is a valid option. To do so, we performed a grid search, where we fine-tuned the task adapter on the basic English dataset for different hyperparameter sets (see Table 4).

learning rates	**epochs**	**batch size**
1e-3, 1e-4, 1e-5	10, 50, 100	8, 16, 32

Table 4: Hyperparameter values used for the grid search for MAD-X. We ran each combination, with a total of 27 hyperparameter sets.

We then used the English dataset by Mohammad et al. (2016) as our validation dataset, and pretended that the datasets for German, Russian and Latin were also validation datasets. We obtained the F1-scores for each hyperparameter set across all four validation datasets (see Appendix A). We then calculated Spearman's rank-order correlation coefficient ρ between the F1-scores for the English validation dataset and the target-language validation datasets. I.e., we examined whether we find a correlation between the hyperparameter sets that lead to high (low) results for the English validation set and the hyperparameter sets that lead to high (low) results for each of the target-language datasets. If the same sets lead to high (low) F1-scores for English and some target language, then we could infer that fine-tuning the hyperparameters on a source-language dataset is sufficient and no target-language material is necessary for the validation. We however found no strong correlation between English and any target language, see top row in Figure 2.

What we did observe, though, was a strong correlation between the target language datasets, which indicates that a dataset from a language other than the source or target language, i.e. from a third language, can be used for validation. Accordingly, we

[13] We only used one seed (42) to produce the results, because our aim is to show variance across shots, not seeds.

[14] Hyperparameter-tuning was carried out for the task adapter, the language adapter was taken off-the-shelf from AdapterHub, see Pfeiffer et al., 2020a.

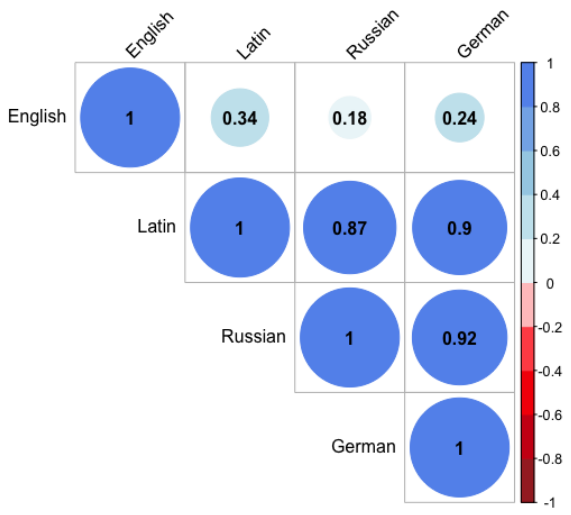

Figure 2: Spearman's rank-order correlations ρ between the hyperparameter sets of the three target languages with regard to the achieved F1-scores for MAD-X.

used the Russian dataset as a validation dataset for the target languages German and Latin (batch-size: 32, learning rate: 1e-3, 50 training epochs), and the German dataset as a validation dataset for Russian (batch-size: 32, learning rate: 1e-3, 100 training epochs).[15] The results are presented in Table 5. Russian shows a result that is comparable to the best results of the other classifiers (except MAD-X with default hyperparameters and the augmented training dataset). The results for German and Latin, in contrast, are the highest across all experiments, and SD is rather low ($<$ 2.5 F1-points).

	ru	ge	la
MAD-X	82.7±2.5	77.3±0.4	73.8±0.9

Table 5: Mean F1-scores (± SD) for using the best performing hyperparameter set from Russian validation data for Latin and German, and the best performing hyperparameter set from German validation data for Russian with MAD-X across three different seeds.

5.4 Summary of Results

MAD-X showed the best performance. For Russian, using default hyperparameters and an augmented training dataset led to the best performance across all models, whereas for German and for Latin hyperparameter-tuning with the basic training dataset led to the best results across all models. These two scenarios (i.e. augmented training dataset, hyperparameter-tuning) also show a small SD across different seeds, which means that the results are robust in terms of different hardware. The results that we obtained with hyperparameter-tuning were generated by using data from a third language (i.e. neither from the source nor from the target language) as validation data. The use of a third language dataset for validation should be confirmed by more experiments for other high- and low-level tasks, as well as for other languages.

When using the basic training dataset (which covers very few training instances) and default hyperparameters, mB0 performed best (only for Russian mB20 showed slightly higher results). mB0 was even able to produce significantly[16] better results than the baseline for Latin, which no other model achieved besides MAD-X. Even though mB20 achieved high results for Russian when using the basic training dataset, all other results are worse or only slightly better than the baseline. As the SD across different shots is very high (see Table 3), it is important to select an appropriate shot. This is inconvenient in a cross-lingual setup, since no validation data in the target language is available. Finding a solution for this problem would be beneficial, since the best shot for German led to results even higher than the results from MAD-X.

Overall, we reach an F1-score of 86.2 for Russian, comparable to Tsvetkov et al. (2014) with an F1-score of 86.0, but the random forest classifier was not able to outperform the neural models.

6 Qualitative Analysis

It was expected that the models perform better on German than on Russian. Afterall, more German than Russian data was used to pretrain mBERT, and German is typologically closer to the source language English than Russian. This expectation was not confirmed. Therefore, we carried out a qualitative analysis. Here, possible sources of errors were identified for German by looking at the predictions of zero-shot mBERT with default hyperparameters and basic training dataset. One hypothesis as to why the models performed worse for German than for Russian is that the target words consist of "computationally challenging" particle verbs (Köper and Schulte im Walde, 2016), i.e. combinations of a base verb (e.g. "schminken") with a prefix parti-

[15] We also applied this hyperparameter-tuning to the other neural and non-neural models, but observed no improvement.

[16] According to χ^2-testing for the model with seed 42 and p<0.05.

cle (e.g. “ab-”)[17]. They are highly productive and notoriously ambiguous. Also, the particle may be separated from the base verb. In contrast, the target words in the Russian dataset are frequent verbs.

Another hypothesis as to why the models performed worse on the German dataset than on the Russian dataset is that the German dataset contains many idioms. For example:

(12) Da wird der Teufel mit dem Beelzebub ausgetrieben. (translation: “One evil is replaced by another.”)

Interestingly, similar variants of this idiom were classified inconsistently. While the target word in (12) was misclassified as literal, it was correctly classified as metaphorical in (13):

(13) Denn die Elite und die USA werden den Teufel nicht mit einem Beelzebub austreiben. (translation: “For the elites and the U.S. will not replace one evil with another.”)

In total, three out of seven sentences that contain the idiom “den Teufel mit dem Beelzebub austreiben” were classified incorrectly. Similar behaviour was also observed for other highly conventionalized expressions, such as “Dampf ablassen” (translation: "let off steam"). In order to test whether the classifier indeed struggles with idioms, the dataset from Ehren et al. (2020) was used. This dataset consists of sentences from 34 preselected verbal idioms. For each sentence the information is given whether it contains a figuratively used idiom or not. In order to make it comparable to our version of the dataset by Köper and Schulte im Walde (2016), it was balanced and reduced to 2000 instances.

All neural models were applied to this dataset. As can be seen in Table 6, the results for the dataset by Ehren et al. (2020) were lower than the results for the dataset by Köper and Schulte im Walde (2016) across all models. This suggests that the neural methods for word-based MD do not work as well on idioms as they do on less conventionalized metaphors, especially since the target words (non-complex German verbs) are less computationally challenging in this dataset than the particle verbs in Köper and Schulte im Walde (2016).

A third hypothesis attributes classifier weakness to instances where the target verb is part of an extended metaphor:

(14) In der Gerüchteküche wurde tagelang deftig aufgekocht. (translation: “For days the gossip factory was working overtime.”)

Here, not only the target word is used metaphorically, but also most context words. This and comparable sentences were misclassified; apparently, too little evidence hinted at the metaphoricity. From 1792 sentences in the balanced dataset (Köper and Schulte im Walde, 2016) that we used for our experiments, 398 were misclassified. We analysed all 398 misclassifications. Our possible explanations regarding idiomatic rather than metaphorical expressions, and regarding larger metaphorical contexts, however, only account for roughly 26 misclassifications. We conclude that the vast majority of instances were misclassified either due to the structural difficulty of particle verbs, or that further reasons for the misclassifications still have to be identified. Additionally, the sentences in the Russian dataset are shorter, which makes it easier for the neural models to make correct predictions: the sentences in the Russian dataset contain an average of nine tokens, while the average sentence length for the German dataset is 13 tokens.

	Ehren	**Köper**
baseline	0.67	0.67
mB0	69.7±3.3	72.5±5.2
mB20	66.9±0.7	70.9±3.9
MAD-X	67.9±2.2	75.2±0.3

Table 6: Mean F1-scores (± SD) for detecting metaphorical usage in the dataset by Ehren et al. (2020) using three seeds (default hyperparameters, augmented training set); results for dataset by Köper and Schulte im Walde (2016) shown in gray.

7 Conclusion

While research on MD has focused on languages with comparably large amounts of data used for pretraining large language models, our experiments have shown that neural cross-lingual methods are suitable for languages with relatively large (Russian and German) and small amounts of pretraining data (Latin). Especially MAD-X performed very well, with the highest results across all experiments for German and Latin using a small training dataset

[17]The literal translation of the particle verb “abschminken” is “to remove makeup”.

and tuned hyperparameters, and for Russian using a large training dataset and default hyperparameters.

Zero-shot classification with mBERT performed decently on a small training dataset and default hyperparameters across all three languages. Few-shot classification with mBERT as applied in our experiments was not successful, as it relies on validation data for shot-selection, which is not possible in the cross-lingual setup. The non-neural random forest classifier, even though it yielded competitive results for Russian and German, was generally outperformed by the neural models – even for Latin, where small amounts of data were used to pretrain the neural models. It is unclear, however, why performance was better for Russian than for German across experiments. A qualitative analysis revealed a range of possible explanations, namely the inherent difficulty of particle verbs, idioms, and rich metaphorical contexts in the German dataset.

Whereas for the few-shot experiments we conducted sequential fine-tuning on source and target language data, Schmidt et al. (2022) showed that joint (instead of sequential) fine-tuning leads to few-shot models that yield higher results and are more robust in terms of hyperparameters (e.g. number of training epochs). We plan to employ this method for MD in future work, because few-shot fine-tuning showed promising results but still depends on target-language validation data. Another next step will be to compare our models' performance for Latin to their performance for Romance languages, in order to minimize the typological differences between the target languages. We will also investigate how the models presented in this paper perform in contrast to newer multilingual large language models such as mT5[18] (Xue et al., 2021).

8 Limitations

The MD methods described in this paper were investigated only for individual, curated sentences. Optimally, however, MD should be carried out on the basis of longer sequences from authentic data; here, also sequence-based metaphor detection should be applied to detect entire metaphorical phrases.

The target languages chosen for the experiments only cover a small subset of languages that were used to pretrain large language models; they should be repeated for other target languages with low amounts of pretraining data, especially those that do not belong to the Indo-European language family. Finally, English is studied as the only source language for the cross-lingual transfer, but it is possible that other languages with rather large amounts of pretraining data might be better suited as source languages.

[18] mT5 was pretrained on modern languages as well as on Latin data (Xue et al., 2021).

9 Acknowledgements

We thank the reviewers for their valuable feedback. We are also very grateful for the support of Filip Miletić, Prisca Piccirilli, Annerose Eichel and Andrea Horbach in the various stages of this study. This research was supported by the DFG Research Grant SCHU 2580/4-1 *(MUDCAT -– Multimodal Dimensions and Computational Applications of Abstractness).*

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.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3036–3045, 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.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220, Online. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, USA.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930, Copenhagen, Denmark. 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.

Zoltan Kövecses. 2010. *Metaphor: A Practical Introduction*. Oxford University Press.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

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.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. Improvements and extensions on metaphor detection. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 33–42, Online. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

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

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association.

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.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. 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.

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.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Joshua R. Westbury, Kris Lyle, Jimmy Parks, and Jeremy Thompson. 2016. *The Lexham Figurative Language of the Bible Glossary*. Lexham Press.

Michael Wilson. 1997. MRC psycholinguistic database: Machine usable dictionary, version 2.00. *Behaviour Research Methods*, 20.

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.

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.

A Hyperparamter-Tuning for MAD-X: Additional material

Table 7 reports the sets of hyperparameters that were used during hyperparameter search for the MAD-X classifier. Figure 3 shows which hyperparameter set led to which F1-score for each of the four languages. This figure hints at the fact that the correlations between English and each of the three languages Russian, German and Latin are low, while the correlation for language pairs not including English are high. We quantified this assumption by calculating Spearman's rank-order correlations presented in Figure 2 (see Section 5.3).

index	learning rate	epochs	train batch size
1	1e-3	10	8
2	1e-3	10	16
3	1e-3	10	32
4	1e-3	50	8
5	1e-3	50	16
6	1e-3	50	32
7	1e-3	100	8
8	1e-3	100	16
9	1e-3	100	32
10	1e-4	10	8
11	1e-4	10	16
12	1e-4	10	32
13	1e-4	50	8
14	1e-4	50	16
15	1e-4	50	32
16	1e-4	100	8
17	1e-4	100	16
18	1e-4	100	32
19	1e-5	10	8
20	1e-5	10	16
21	1e-5	10	32
22	1e-5	50	8
23	1e-5	50	16
24	1e-5	50	32
25	1e-5	100	8
26	1e-5	100	16
27	1e-5	100	32

Table 7: Index to hyperparameter sets for MAD-X.

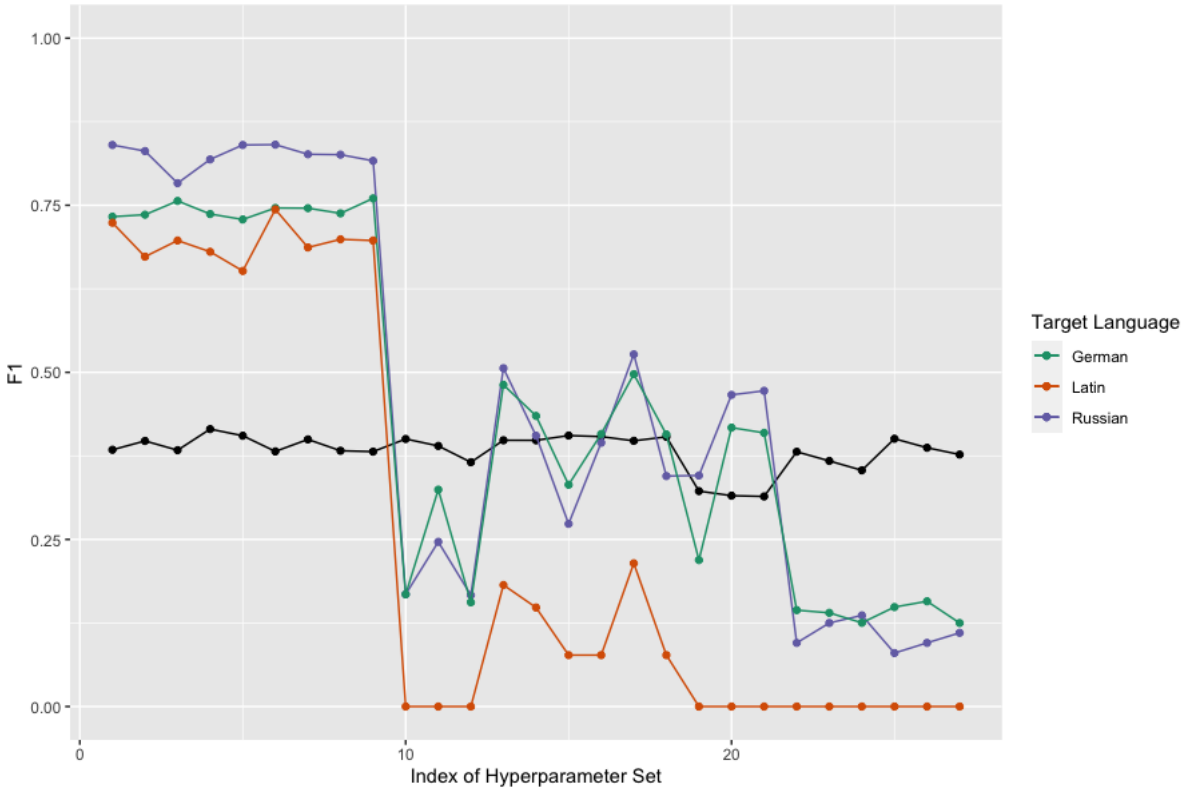

Figure 3: Result for using both the data from Mohammad et al. (2016) (black line) and the different test sets for target languages Russian, German and Latin as dev sets for the grid search on zero-shot classification with MAD-X.