Applying Transfer Learning to German Metaphor Prediction

Maria Berger, Nieke Kiwitt, Sebastian Reimann
Ruhr University Bochum
{maria.berger-a2l, firstname.lastname}@rub.de

Abstract

This paper presents results in transfer-learning metaphor recognition in German. Starting from an English language corpus annotated for metaphor at the sentence level, and its machine-translation to German, we annotate 1000 sentences of the German part to use it as a Gold standard for two different metaphor prediction setups: i) a sequence labeling set-up (on the token-level), and ii) a classification (based on sentences) setup. We test two transfer leaning approaches: i) a group of transformer models, and ii) a technique that utilizes bilingual embeddings together with an RNN classifier. We find out that the transformer models do moderately in a zero-shot scenario (up to 61% F1 for classification) and the embeddings approaches do not even beat the guessing baseline (36% F1 for classification). We use our Gold data to fine-tune the classification tasks on target-language data achieving up to 90% F1 with both, the multilingual BERT and the bilingual embeddings. We also publish the annotated bilingual corpus.

Keywords: Information Extraction, Information Retrieval, Language Modelling, Less-Resourced/Endangered Languages, Neural Language Representation Models, Multilinguality, Word Sense Disambiguation

1. Introduction

Being a part of figurative language detection, metaphor recognition is one of the most challenging tasks in natural language processing. However, especially in languages other than English—and even there—training and testing resources for automated tasks are rare.

There exist multiple ways to address automated metaphor recognition. Earlier techniques often work based on noun-verb pattern extraction in text (Shutova, 2010; Tsvetkov et al., 2013), or are dictionary-informed (Steen et al., 2010; Tsvetkov et al., 2014, 2013). Other linguistically-informed approaches make use of linguistic features (such as abstractness scores) together with some statistical measures to derive metaphoricity (Turney et al., 2011; Shutova et al., 2013; Shutova, 2010; Köper and im Walde, 2017), or use even early variants of nowadays wide-spread embeddings representations (Latent Semantic Analysis) (Kintsch, 2000).

Neural approaches are now rapidly ascending. Often, different architectures of artificial neural networks are used to process the semantic representation of words within texts (GloVe, Word2Vec) to find metaphors at the word level (Maudslay et al., 2020; Gao et al., 2018; Bizzoni and Lappin, 2017; Bizzoni and Ghanimifard, 2018). Other neural approaches include semantic information derived from FrameNet (Li et al., 2023) or a target-oriented parse tree structure together with a RoBERTa-based transformer model (Wang et al., 2023). Finally, Choi et al. (2021) consider contextualized representation (delivered by pre-trained transformer models), and also consult with linguistic metaphor identification theories such as Selectional Preference Violation (Wilks, 1975) and the MIP guidelines (Group, 2007) that identify the gap between the contextual and literal meaning of a word, to decide whether a target is metaphorical.

Even though research work in automated metaphor detection especially increased in recent years, there exists only a comparatively low number of works that investigate metaphor recognition in languages other than English (Sanchez-Bayona and Agerri, 2022; Aghazadeh et al., 2022).

We think that, due to the conceptual nature of metaphor (Lakoff and Johnson, 1980), it is sufficient to expect a successful language transfer for metaphoric expressions given a sufficient amount of data that is capable to encode this conceptual nature.

The research question that drives us in the first place is **RQ**: (How well) are existing transfer-learning models applicable to metaphor prediction in German (spoken) language text.

In this article, we perform a study to assess state-of-the-art technology—namely multilingual BERT, XML-RoBERTa as well as a multilingual embeddings setups—for cross-language metaphor recognition in a translated German metaphor data set. Precisely, we train (in case of LLMs fine-tune) a metaphor recognition system based on an English language corpus, and then apply it to the German part of another corpus. The German metaphor data we are using comprises the first 1000 sentences of a machine-translated, manually metaphor-annotated extension of the corpus by Gordon et al. (2015). We publish our data together with this article.[1]

[1] https://doi.org/10.6084/m9.figshare.

LREC-COLING 2024, pages 1383–1392
20-25 May, 2024.

2. Related Work

2.1. Neural Metaphor Prediction in Languages other than English

Successful work in cross-lingual automated metaphor detection was performed by (Tsvetkov et al., 2013, 2014). The authors use lexical-semantic word features (which can contribute to metaphorical construction) and bilingual dictionaries (for Spanish, Farsi, Russian) as a data base for transfer learning that recognizes metaphorical expressions across languages. In a comparable setup, Tsvetkov et al. (2013) are using syntactic patterns, WordNet's semantic categories and a vector representations-based abstractness score (MRC Psycholinguistic Database) in their work. The authors trained a classifier on English samples and applied it to a target language. The method also obtains semantic features from the target language. The authors hypothesize that when either subject or object of a concrete verb is abstract (becoming inconsistent with the verb) then the verb might be used figuratively.

Aghazadeh et al. (2022) apply a probing mechanism in metaphor-annotated data. Utilizing pretrained multilingual language models, they probe for cross-lingual performance in a data set of four high-resource languages (English, Russian, Spanish, Farsi).

Berger (2022) investigate vocabulary coverage of embeddings trained on Europarl. First results, however, showed low performance, because, amongst others, the authors did not consider stop words during training the embeddings.

2.2. Non-English Datasets

Sanchez-Bayona and Agerri (2022) publish a Spanish language, multi-domain metaphor corpus, which is sufficient for neural metaphor detection tasks. The authors also follow MIPVU (Steen et al., 2010) metaphor annotation guidelines. The work presents one of the rare cases of transferably learned metaphor for languages other than English.

Lu and Wang (2017) annotate a corpus of Mandarin Chinese according to the MIPVU guidelines. While evaluating MIPVU's transferability for Chinese, the authors also investigate the proportion of metaphor-related words across different registers (academic, fiction, news) in Chinese texts. The resulting corpus totals 30,000 words.

Our work gives a first overview of different transfer learning strategies for metaphor prediction from English to German language without any explicitly engineered linguistic features. It also is—to the best of our knowledge—the first work that investigates transfer learning from English to German metaphors, beyond multilingual transformers and beyond sequence labeling only.

3. SEQUENCE LABELING

We experiment with two different setups for sequence labeling. First, we test three different transformer models, and second, we apply the sequence labeling technique by Gao et al. (2018) using bilingual embeddings data. Gao et al. (2018) use GloVe (Pennington et al., 2014) and ELMo embeddings (Peters et al., 2018) data as input representations. We instead use bilingual embeddings to perform cross-lingual metaphor prediction. Our experiments in sequence labeling use the following models:

1. Transformers: mBERT (Devlin et al., 2018), XLM-RoBERTa (Liu et al., 2019), sentence transformers (SBERT) (Reimers and Gurevych, 2019)

2. An RNN architecture that makes use of bilingual embeddings (e.g., Europarl by Koehn (2005), CommonCrawl by EMNLP (2018), and News Commentary data[2])

3.1. Training and Testing Data

Throughout all our experiments, we use the English VUA corpus (Steen et al., 2010) as training data (for LLMs fine-tuning), and a German corpus that we expanded from the English language metaphor corpus by Gordon et al. (2015)[3] as testing data. Precisely, we use Google translate to produce a German version of the data. We manually evaluated the quality on a 3-scale rating.[4] A smaller part of the corpus by Gordon et al. (2015) was translated into German and manually evaluated before (Berger, 2022).

Following the splits introduced by Gao et al. (2018) the training, validation and testing data sizes of the VUA corpus are 6,323, 1,550 and 2,694 sentences respectively. For testing in German, two authors manually labeled 98 sentences

25047683.v1

[2] https://www.statmt.org/wmt13/translation-task.html, accessed: Jan 2024

[3] This effort contains or makes use of the IARPA-funded Metaphor Program USC/ISI annotated metaphorical language collection, release iarpa_metaphor_isi.edu_metaphor_corpus_2015-04-03. https://figshare.com/articles/dataset/A_Corpus_of_Rich_Metaphor_Annotation/6179210, accessed: Jan 2024

[4] 842 of the sentences are rated as high (appropriate translation), 113 as mid (one falsely translated stop word), and 45 as low (broken translation)

that were randomly selected from our German metaphor corpus (see Tab. 2 for all numbers) while our third co-author consulted in the process. Below, we show an example of a token-labeled sentences where 1 represents metaphoric expression, and 0 represents literal meaning.

(1) *Ich_0 werde_0 heute_0 draußen_0 in_0 der_0 Stadt_0 sein_0 und_0 den_0 weinigen_1 Blutstrom_1 spüren_1 ,_0 den_0 apfelroten_1 Kreislauf_1 der_0 Demokratie_0 ,_0 ihr_0 fleischliches_1 Wissen_0 ohne_0 Weisheit_0 ._0*

[I will be out in the city today, feeling the vinous veinous thrust of blood, the apple-red circulation of democracy, its carnal knowledge without wisdom.]

For the labeling procedure, we follow the MIPVU guidelines (Steen et al., 2010). MIPVU requires the annotator to, for each word, identify the meaning of the word in the context it is used, then look it up in the dictionary and determine if a "more basic" (that means more concrete or human-oriented) meaning is present. If such a more basic meaning can be found, then, as a last step, decide if these two meanings are related by similarity and still sufficiently distinct. If both points are true, then the word can be labelled as metaphor. These guidelines offer a tremendous help during the decision process on whether a word is used metaphorically or literally in the given context.

3.2. Transformers

Metaphor prediction is usually modeled as a sequence labeling problem (Bizzoni and Ghanimifard, 2018; Mao et al., 2019; Gao et al., 2018). Hence, we make use of three different transformer models pre-trained for token classification. The utmost used multilingual BERT model (Devlin et al., 2018)[5], second, XLM-RoBERTa (Liu et al., 2019), third, a smaller sentence transformer model (SBERT) (Reimers and Gurevych, 2019), which is faster—even though a bit less performing—and was trained using a Siamese network approach. We fine-tune all models on English metaphor-annotated data for 5 epochs with a batch-size of 16 samples.

4. CLASSIFICATION

In this section, we model metaphor detection as a classification task. We train (in case of LLMs fine-tune) a binary classifier on the VUA verb annotated data set and test on the English and German share of the metaphor corpus. Even though metaphor prediction is typically a sequence labeling problem, in this section, we present results for a classification problem.

Experiments overview:

1. We design a baseline classification technique that is based on a word's rank and a simple neighborhood measure.
2. We apply a variant of index encoding to transformer models (mBERT, XLM-RoBERTa, SBERT) by encoding each sample twice, once with attention at all positions, once with attention at the specified position only.
3. We apply the approach by Gao et al. (2018) to our testing data utilizing bilingual embeddings (based on Europarl, CommonCrawl, News Commentary) together with an RNN classifier.

4.1. Training and Testing Data

Again, we use the English VUA corpus for training data, and the German metaphor corpus for testing. We use the training, validation and testing data splits from the VUA corpus (train: 15,516; val: 1,724; test: 5,873 accordingly). To compare transfer to German, we also consider testing performance on the VUA data.

Our testing data comes together as follows.

English testing data: We use the original corpus by Gordon et al. (2015) (English language part), which consists in 1789 sentences annotated for metaphor source, metaphor target, lexical trigger and some categorical information of the linguistic metaphors. For the positive class, we retrieve those sentences in which the metaphoric expression is annotated as verb (this happens 1016 times). For the negative class, we use the remaining sentence, and make sure that there is a verb tagged that is not a metaphor (this happens 685). Summarizing, this data part contains 1701 instances (M-En), because 88 sentences do not contain any verb.

German testing data: From the 1000 annotated sentences of the German corpus share, we also select those sentences with the annotated metaphor being a verb and put them into the positive class (this happens 521 times). For the negative class, we consider the remaining sentences and search for a verb that is not labeled as a metaphor (this happens 387 times). Summed up, this data part contains 908 instances (henceforth, we call it "Met-DE" for German Metaphor data), because 55 sentences do not contain any verb and 37 sentences more do not contain a metaphor (we

[5] https://github.com/google-research/bert/blob/master/multilingual.md, accessed: Jan 2024

embed.	test (samples)	tokens	precision	recall	F1-score	accuracy
Europarl	Met-EN (1016:685)	1,544	64	65	65	57
	Met-DE (521:387)	875	69	49	58	58
Common Crawl	Met-EN (1016:685)	1,585	67	70	**69**	61
	Met-DE (521:387)	877	71	52	**60**	60
News Commentary	Met-EN (1016:685)	1,482	62	62	62	55
	Met-DE (521:387)	831	68	50	58	58

Table 1: Results (%) of the unsupervised **classification baseline** (rank r$<$ 25% and #neighbors $>$ 10); *samples* declare the number of candidates in each class; *tokens* represents the in-vocabulary share of the sample sizes

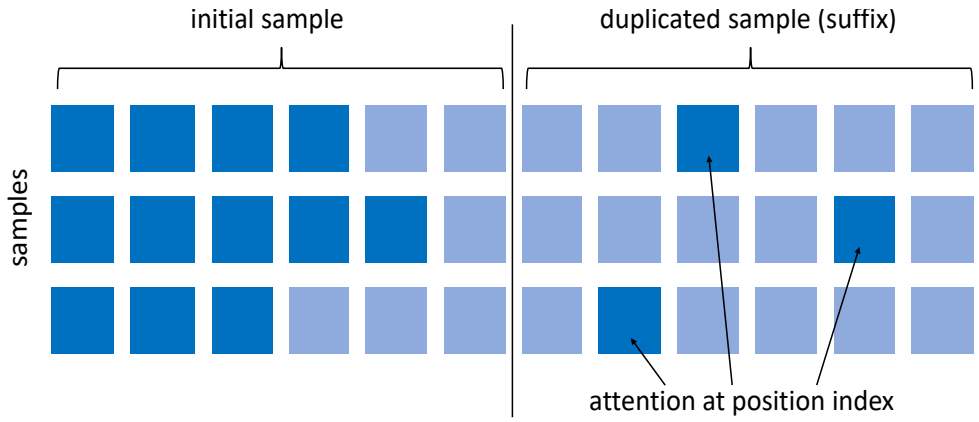


Figure 1: Illustration of the index attention for the classification set-up of metaphor prediction; the attention is 1 for tokens, 0 for margin; in the duplicate, attention is 1 at the metaphor index, 0 in the remaining sentence

skipped them upfront already). We refer to this dataset by Met-DE henceforth.

4.2. Baseline

Del Tredici and Bel (2016) investigate the relation between metaphoricity and the distribution of a (potentially metaphoric) verb by considering a corpus-specific measure. The authors start from the conjecture that some verbs are used in contexts of highly metaphoric expressions, while others are not. Clustering the verbs' contexts and frequencies, they find a frequency measure that they successfully test for metaphor recognition. We strongly simplify the approach by Del Tredici and Bel (2016), to design a primitive baseline technique. Precisely, we consider only VERB and NOUN[6] and introduce two parameter thresholds:

- $number\_of\_neighbors > 10$, and
- $rank < 25\%$

within a window of 5 of a given word in the respective context. If both conditions are true, the word is labeled as a metaphoric word. Table 1 shows that this simple technique works already well for the Met-DE and the M-En data, and that it is rather robust against the size of parallel data. This becomes especially clear looking at the Met-DE F1-scores using CommonCrawl embeddings (F1 60%) and News Commentary embeddings (F1 58%) while CommonCrawl contains over 2 m. lines and the News Commentary corpus contains about 180,000 lines.

4.3. Transformers

Metaphor prediction is typically a sequence labeling problem, but we also model the metaphor prediction task as a classification task. We do this because our initial corpus only contains one annotated metaphor per sentence. To make use of that information, it is critical to assess how a typical sequence labeling problem can be represented as a classification problem. Gao et al. (2018) approach the problem by encoding a suffix embeddings representation at the end of each embedded sample. This means, they append another embedding representation at the end of the vector representations of their samples (sentences), but only "unmasking" the word of interest while the remaining words are encoded with 0. This way, with the input training one can hand over a position index to each sample, and then, only this position is encoded as a suffix.

We attempt to transfer this idea to the attention

[6]The vast majority of our metaphors are verbs and nouns even though some adjectives are possible too.

approach in transformer models. We duplicate the input sample while setting attention only to the specified index (of the potentially metaphor) in the duplicate. This is done in transformers by masking the other tokens with 0 ad only allowing attention of 1 for the metaphoric word. Figure 1 illustrates this procedure.

4.4. Bilingual Embeddings and RNN

We train bilingual embeddings based on three different parallel corpora. We merge the parallel data of each corpus by a simple zip-like merging strategy. We do not remove stop words during embeddings training.[7] Following parallel data present the base of the bilingual embeddings:

- The English/German part of Europarl Parallel Corpus (Europarl)(about 1.9 m lines)[8]
- The Common Crawl corpus (EMNLP, 2018) (about 2.4 mio. lines), and
- The News Commentaries from the machine translation challenge[9] (about 180,000 lines)

Afterwards, we apply an RNN architecture by Gao et al. (2018) using input representation from the trained bilingual embeddings.[10].

5. Results & Evaluation

We split our results section into two parts according to the sequence labeling and the classification results and techniques.

5.1. Sequence Labeling

Transformers Approaches: To gain a first understanding of transformer performance, we first show recall and precision numbers of all three transformer models used (c.f., Figure 2). At first glance we see that SBERT behaves a bit different than mBERT and RoBERTa in the smaller part-of-speech classes (less than 50 metaphoric contexts in the testing data). First, SBERT's recall of metaphor adjectives is much higher, but, at the same time, its precision in predicting them is also much lower, which basically compensates the former. Further, SBERT's recall of adpositions is also much lower compared to mBERT and RoBERTa. One explanation is, considering that SBERT is significantly smaller than the other two models, it is possible that it can not handle the minor classes very well, as it tends to overgeneralize. We can see that for the two major classes (VERB and NOUN) SBERT behaves very similar to the other two models. Another peculiarity is that XLM-RoBERTa does not find any adverbials. The dynamic masking process during training RoBERTa initially is much more flexible compared to the static masking of BERT. One advantage is that RoBERTa is better in generalizing to new data points. However, in our semantically more challenging set-up, this flexibly might prevents RoBERTa from retrieving rather unknown items.

As shown in Table 2, we find a comparably low precision for Met-DE using mBERT and SBERT (30% and 23%), while recall is 7% points higher. We look at samples from the first 9 sentences to obtain an understanding of the labeling problems from a closed-reading perspective. We choose the output of mBERT[11].

The first sample is also the example in Sec. 3. The model predicted only one token falsely *Weisheit(FP) (wisdom)*, which might can be explained by its rareness.

We further find false positives in:

(2) ***stetigen**(FP) Strom **von**(FP)*
[**steady** flow **of**]

(3) ***bis**(FP) **zum**(FP) Äußersten **erhöhte**(FP)*
[**increased to** extremes]

(4) ***In**(FP) den 1760er Jahren **gab**(FP) es eine viel **größere**(FP) Bedrohung*
[Back **in** the 1760's there was a far **greater** amount of threat]

Example (2), (3) and (4) contain prepositions, which often are part of metaphoric expression in the English training data.

(5) *Einige würden Sie glauben **machen**(FP), dass **diese**(FP) Gesetzgebung den Identitätsbetrug der Wähler heilt, **fuhr**(FP) Wilson **fort**(FP).*
[Some would **have** you believe that **this** legislation cures voter identity fraud, **continued** Wilson.]

Here, possibly the German syntax has a confusing effect on *machen* as well as the split of *fuhr fort*. Further, *fortfahren* indeed is an ambiguous word, even though it is not a metaphor, because—in our

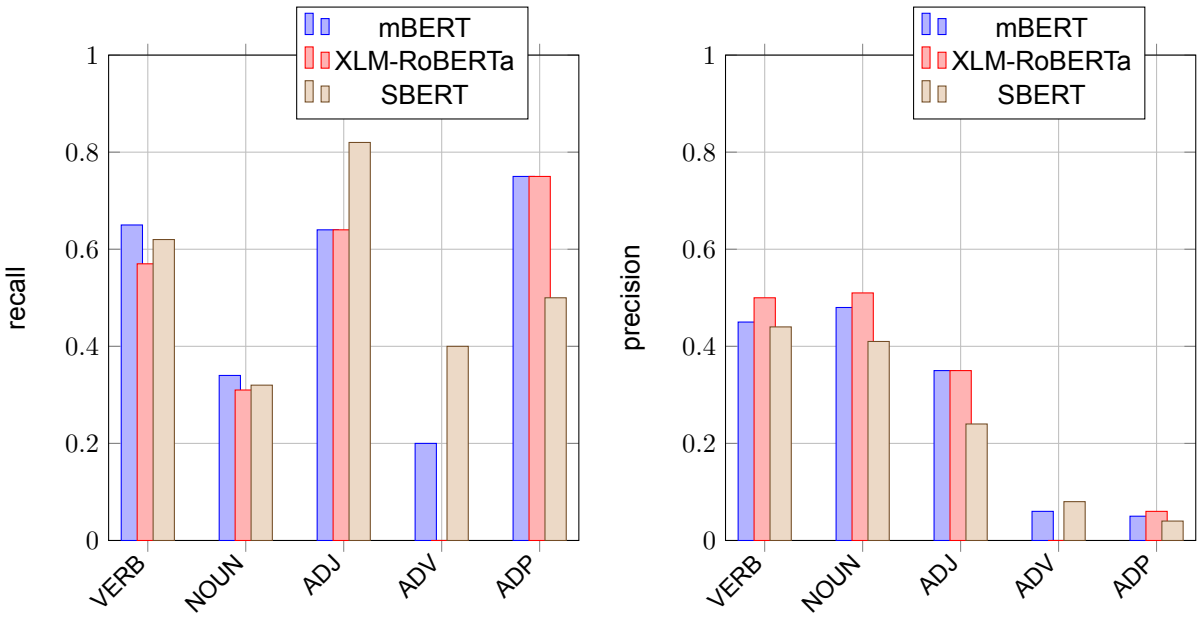
[7] We use word2vec (Mikolov et al., 2013) python package for training the bilingual embeddings with vector length 300, a min frequency of 5 and a window of 5.

[8] https://www.statmt.org/europarl/, accessed: Jan 2024

[9] We manually re-aligned this-one as it was tremendously out of sync.

[10] We train the RNN for 20 epochs; with a learning rate of .005, Adam optimization and a dropout of .5 and .1 from the input and towards the output respectively; 300 embedding dimensions

[11] XLM-RoBERTa uses a different tokenizing technique



(a) Recall performance of transformer models

(b) Precision performance of transformer models

Figure 2: Recall and precision figures of the transformer models for the zero-shot cross-lingual sequence labeling task by POS-tag

representation	test data (sample size)	test tokens	precision	recall	f1-score	accuracy
mBERT	VUA	62,886	79	71	75	94
XLM-RoBERTa	(2694)	50,177	81	71	**76**	94
SBERT		62,886	74	66	70	93
mBERT	Met-DE	2,316	27	49	**35**	86
XLM-RoBERTa	(98)	2,550	30	44	**35**	87
SBERT		2,316	25	47	33	85

Table 2: Results (%) of **sequence labeling** using different transformer models; trained on VUA corpus with train-val split of 6,323:1,550;

understanding—both meanings are too far apart from each other.

(6) ***macht****(FP) jeden Tag sonnig*
[**makes** every day sunny]

Here, possibly the high frequency and the strong ambiguous character of *machen (make)* causes problems to the classifier, especially since *make* was frequently annotated as metaphoric in the VUA training data.

(7) *die Feindseligkeiten* ***zwischen****(FP) den Nationen für immer* ***aufhören****(FP)*
[hostilities **between** nations **cease** forever]

We do not have an explanation for these cases. However, we found that in the VUA corpus, especially stop words such as “after”, “on” or “in” are annotated as metaphorically in certain context, which can affect predictions of FP in the German data.

(8) ***von****(FP)* ***enormer****(FP) Verantwortung*
[**from tremendous** responsibility]

We suppose that the preposition together with the English language source word has an affect on the falsely labeling here as well, because tremendous has a stronger metaphoric tendency than the German pendent.

Concerning false negatives (FN), we annotated:

(9) *[...] die(FN) seelischen(FN) Wunden des(FN) Krieges(FN) heilt*
[Democracy heals the mental scars of war]

completely positive for metaphoric expression, because we understand that phrase as a personification with respect to “Democracy”. However, only *Wunden* and *heilt* was recognized by the system. In fact, this is an example for uncertainty in the annotation process, because indeed only “Wunden” and “heilt” are taken from another domain while *seelischen* and other words from that phrase are not necessarily.

In contrast, in

(10) *Symptome(FN) heilen*

representation	test data (sample size)	precision	recall	f1-score	accuracy
mBERT	VUA	73	63	**68**	82
XLM-RoBERTa	(5,873)	61	44	51	75
SBERT		65	54	59	78
mBERT	Met-EN	64	57	60	55
XLM-RoBERTa	(1,701)	60	54	57	51
SBERT		66	65	**66**	59
mBERT	Met-DE	**58**	46	52	50
XLM-RoBERTa	(908)	58	44	50	50
Sentence BERT		57	65	**61**	51
fine-tuned in TL					
mBERT	Met-DE	**91**	88	**90**	88
XLM-RoBERTa	(98)	81	86	83	81
SBERT		73	**92**	82	78

Table 3: Results (%) of the **classification** task using different transformer models; upper part: trained on VUA corpus with train-val split of 15,516:1,724; lower part: fine-tuned using VUA corpus of 15,516 samples and Met-DE train-val spilt of 720:90

embeddings	voc addr	voc size	test data	p	r	f1	ac
Europarl	11,356	18,695	VUA	65	54	59	77
CommonCrawl	12,976		(5,873)	67	55	**60**	78
News Commentary	9,556			63	39	48	75
Europarl	12,134	18,992	Met-EN	67	57	62	58
CommonCrawl	13,825		(1016:685)	65	63	**64**	58
News Commentary	10,455			62	46	53	51
Europarl	14,079	20,949	Met-DE	73	23	35	51
CommonCrawl	15,527		(521:387)	56	27	**36**	46
News Commentary	12,340			68	10	18	46
fine-tuned in TL							
Europarl	13,987		Met-DE	90	**93**	**91**	90
CommonCrawl	15,403	20,732	(98)	**91**	84	87	86
News Comm.	12,274			80	75	77	76

Table 4: Results (%) of the **classification** task using different embeddings representations considering stop words and lower-cased data; voc(ab) addr(essed); voc(ab) size; upper part: trained on VUA corpus with train-val split of 15,516:1,725; lower part: fine-tuned using VUA corpus of 15,516 samples and Met-DE train-val spilt of 720:90; tested using Met-DE of 98 samples; RNN trained 20 epochs, vectors length=300 test samples present class sizes; RNN trained 20 epochs, vectors length=300

[heal symptoms]

Symptome was not recognized. Possibly, a stronger context for metaphors makes the system more sensitive to it. In

(11) *was die muslimische Welt(FN) wirklich quält*
[what really ails the Muslim world]

however, whether *Welt* is a personification or might be understood literally is not easily to decide. Other difficult examples come from the religious domain. During annotation, we decided that

(12) *Segen(FP) Gottes* [blessing from God]

is not a metaphor, while

(13) *Ungnade(FN) [Gottes]*
[God's disfavor]

is one. However, the system decided conversely.

We can summarize that cross-lingual sequence labeling of metaphoric speech is a challenging task. First, because the syntactic and lexical choices have a huge impact to the model, second, because senses also differ. Further, annotation disagreements between the initial VUA guidelines, and our own perception of metaphors as well as dictionary readings in German dictionaries can impact the results.

5.2. Sentence Classification

Transformer Approaches: As shown in Tab. 3, only the small SBERT model slightly beats our baseline (61% versus 60%). Even for the English language (M-En) test data, the baseline still

achieves 69% while the best transformer (SBERT) reaches only 66%. After fine-tuning, SBERT shows the lowest F1-score (increase: 61% to 82%) while BERT increases from 52% to 90% in F1-score.

SBERT achieves the best F1-score and by far the best recall (65%), hence, for sample evaluation, we look at data from SBERT: We find false negative especially when the verb is at the very last position in a sentence. Such as in: *Das Geld [...] könnte mich eine Woche lang* ***ernähren****. (The money [...] could feed me for a week)*. Other examples of false negatives are ***schmeckt*** *Demokratie ziemlich süß (democracy tastes pretty sweet)* and *die Ambrosia des Reichtums* ***schmeckt*** *gut (the ambrosia of wealth tastes good)*. One explanation of these falsely labeled "literally" verbs is that the verbs occur with other words from the same domain. Hence, the metaphor is more complex, and more difficult to detect.

True positives are recognized correctly when the verb directly followed a noun, as in *Reichtum* ***lindert*** *(Wealth alleviates)* and *Bürokratie* ***heilt*** *(bureaucracy cures)*. We also find that verbs correctly identified for metaphoric use are those that appear more frequent in the corpus such as *heilt (cures)* and *lindert (alleviates)*.

We find that false positives often happen based on our (sometimes false) assumption that when the metaphor labeled in a sentence was not a verb, then any verb in that sentence might rather being used in a literal context. It turns out that actually, often, these verbs are used metaphorical as well. Such as in: *den weinigen Blutstrom* ***spüren*** (see example above)[12] and in *in die Staatskasse* ***fließen*** *(going into the governmental coffers)*. Also, in *Der Finanzmanager* ***erstellt*** *Finanzberichte (The financial Manager prepares financial reports)* and *Der Arabische Frühling* ***entstand*** *(The Arab Spring arose)* verbs are false positives, because they are part of a personification. Still, because our entire data set is not labeled at the token-level, we can not fully work around that yet.

Bilingual Embeddings: As shown in Tab. 4, we encounter a massive performance drop when we apply the bilingual embeddings approach together with the RNN classifier to the German language data share. Neither for the Met-EN nor the Met-DE test data, we can beat our baseline of 69% and 60% in F1-score, instead we only reach 64% and 36%. However, when we look at fine-tuning (see Tab. 4), we can see that the bilingual embeddings approach, especially using Europarl and CommonCrawl, even out performs the multilingual transformers showing an F1 score of 91% and 87% respectively. This shows that mBERT performs well in tasks dependent on morphology and syntax, however, it lacks behind in semantically challenging task where machine translation with a high quality estimate is required (Libovický et al., 2019). Metaphor detection definitely counts towards the latter.

[12]Here, the sequence *Blutstrom spüren* was recognized in the sequence labeling task.

6. Conclusion

We presented a comprehensive study of transfer learning techniques—utilizing transformers and bilingual embeddings—to German metaphor prediction. We developed a German metaphor-annotated corpus as an extension of an English corpus that was machine-translated.

For future work, we plan to investigate differences between an original English-German translated and an original German-English translated corpus concerning lexical and conceptual distribution differences. We also plan to use more linguistically-informed techniques of metaphor retrieval to combine both, prediction of target language metaphor recognition (starting from target language fine-tuning), and language-agnostic metaphor retrieval. As shown in this paper, zero-shot transfer approaches seem to be not very stable for cross-language metaphor recognition, hence, we need to find language-agnostic ways to flag metaphoric expressions in a target language.

7. Ethics Statement

All data sets and techniques used are publicly available and credited and references appropriately in our research. We are aware of possible biases of large language models, that can be susceptible to generate harmful content or influence results (Bender et al., 2021). To keep computing costs associated with the use of large language models low, we reuse models, and code as far as possible. We do not re-train transformer models from scratch, instead we fine-tune them in downstream tasks.

8. Acknowledgment

Sebastian Reimann was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), SFB 1475, Project ID 441126958.

9. Bibliographical 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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?□□. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Maria Berger. 2022. Transfer learning parallel metaphor using bilingual embeddings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 13–23. Association for Computational Linguistics.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and bilstms two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101.

Yuri Bizzoni and Shalom Lappin. 2017. Deep learning of binary and gradient judgements for semantic paraphrase. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

M Choi, S Lee, E Choi, H Park, J Lee, and D Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *NAACL-HLT 2021-2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 1763–1773. Association for Computational Linguistics (ACL).

Marco Del Tredici and Nuria Bel. 2016. Assessing the potential of metaphoricity of verbs using corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4573–4577.

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*.

EMNLP. 2018. Emnlp 2018 third conference on machine translation (wmt18) - shared task: Machine translation of news. https://www.statmt.org/wmt18/translation-task.html. Accessed: Oct. 2022.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Jonathan Gordon, Jerry R Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66.

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

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

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.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert?

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of mandarin chinese. *Language Resources and Evaluation*, 51:663–694.

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.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226. Association for Computational Linguistics.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics 21–4*.

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.

Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409, Dubrovnik, Croatia. Association for Computational Linguistics.

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