Transfer Learning Parallel Metaphor using Bilingual Embeddings

Maria Berger
Ruhr University Bochum
maria.berger-a21@rub.de

Abstract

Automated metaphor detection in languages other than English is highly restricted as training corpora are comparably rare. One way to overcome this problem is transfer learning. This paper gives an overview on transfer learning techniques applied to NLP. We first introduce types of transfer learning, then we present work focusing on: i) transfer learning with cross-lingual embeddings; ii) transfer learning in machine translation; and iii) transfer learning using pre-trained transformer models. The paper is complemented by first experiments that make use of bilingual embeddings generated from different sources of parallel data: We i) present the preparation of a parallel Gold corpus; ii) examine the embeddings spaces to search for metaphoric words cross-lingually; iii) run first experiments in transfer learning German metaphor from English labeled data only. Results show that finding data sources for bilingual embeddings training and the vocabulary covered by these embeddings is critical for learning metaphor cross-lingually.

1 Introduction

In the literature, figurative language is instantiated in many different ways. One of the most challenging tasks of figurative language detection, however, is metaphor identification. Dorst (2015) finds that up to almost 20% of words in a text are metaphor-related. However, most work in the field is focused strongly on the English language.

As such, early work on computational metaphor interpretation was performed by Kintsch. Kintsch (2000) uses **Latent Semantic Analysis** to adjust the meaning of a predicate P when it is applied to an argument A. In Kintsch's theory the predicate is what we typically call a metaphor's source (Lakoff and Johnson, 1980) and the argument is its target (e.g., selfies [target] go viral [source]). Before word embeddings were used based on implementations such as word2vec, LSA helped to generate high-dimensional semantic spaces using singular value decomposition for dimension reduction. Kintsch uses cosine similarity to compare a metaphorical predication (i.e., its numerical representation) to some of its semantic surroundings.

Today, semantic information mainly is encoded by word embeddings (Mikolov et al., 2013). Gao et al. (2018) recently presented work of metaphor prediction using an RNN classifier. Together with different types of embeddings vectors, the authors perform **neural metaphor detection**, in a sequence labeling setup, and in a classification setup.

One of the most famous works regarding the development of training and testing data sets is delivered by Steen et al. (2010). The authors present a method for the identification of metaphor in language at the word-level based on methodological and **empirical corpus-linguistic work** in English and Dutch. The method formulates manual instructions and is a refinement based on the metaphor identification procedure (MIP) presented by (Group, 2007). The extended annotation version (MIPVU) is developed at Vrije Universiteit Amsterdam (VUA) and demonstrates case studies addressing metaphor in English and Dutch news amongst others.

While there is a lot of room for improvement in the field of metaphor detection and interpretation, especially languages other than English lack resources and successful algorithms. Transfer learning (TL) is one way to overcome this issue. But work in this field is rare.

Tsvetkov et al. (2014) use lexical semantic features of words participating metaphoric construction. The authors use **transfer learning** based on bilingual dictionaries to find metaphoric expressions across languages. Their work supports the consensus that metaphors are rather conceptual.

More recently, Aghazadeh et al. (2022) perform **probing of metaphor**-annotated data sets. Next to other tasks, they also probe for cross-lingual

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 13 - 23

performance using a multilingual pre-trained language model and a data set of four high-resource languages (English, Russian, Spanish, Farsi).

In this paper, we present different strategies to overcome resource gaps using transfer learning strategies. We start with a literature overview before we perform first experiments to assess these techniques for the German language.

2 Literature on transfer learning

2.1 Types of transfer learning

TL in general refers to techniques applied across different domains and languages. Cross-language (CL) learning refers specifically to the transfer from one language to another while domain adaption (DA) rather showcases the transfer of a technique from one domain to another within the same language. In their comprehensive survey, Weiss et al. (2016) differentiate (among others) between instance-based and feature-based techniques of TL.

Instance-based transfer: Instance-based TL infers knowledge based on the behaviour of instances in a source versus target domain. As such, it attempts to reduce the marginal distribution difference ($P(X_t) \neq P(X_s)$, e.g., word freq.) by reweighting the samples in the source domain to correct for distribution differences (Asgarian, 2018).

One example for instance-based TL is Asgarian et al. (2018). For training, the authors only use information from relevant re-weighted instances in the target domain. The target samples are selected upfront based on the uncertainty (distance of sample x to the decision boundary) in a binary model trained on source and target samples. Also, Jiang and Zhai (2007) find relations between different instance distributions in source and target. They formulate requirements for instance distribution and classification function different in source and target. Then, they solve for these differences using semi-supervised instance-weighting. Dai et al. (2008) migrate knowledge—from labeled data—from a source feature space to a target feature space. The authors show that one can use for example labeled text data to train a model for image classification when image labels are rare.

Feature-based transfer: Feature-based TL aims to reduce the gap between the marginal ($P(X_t) \neq P(X_s)$, e.g., word frequencies) and conditional distributions ($P(Y_t|X_t) \neq P(Y_s|X_s)$, typically Y-labels) of source and target domain (Long et al., 2013). In asymmetric feature-based TL, often a transformation ϕ_s/ϕ_t from source to target is employed (Long et al., 2013), which especially works well when both domains share the same label spaces. In symmetric feature based TL, features are transferred from source and target respectively into a common space. Pan et al. (2010) transfer components across domains into a reproducing kernel Hilbert Space using maximum mean discrepancy as a distance measure. In the sub-space represented by that Hilbert Space, data properties are preserved and data distributions of different domains can still remain similar. This enables the training of classifiers in a source domain for use in a target domain. Also Duan et al. (2012) consider the use of source domain and target domain data represented by heterogeneous features of different dimensions. Two projection matrices help to transform data from source and target into a common subspace, and two feature mapping functions use these projections to augment the data in that new space.

2.2 Task-oriented techniques

Following, we give an overview on TL techniques from a more task-driven perspective.

Cross-lingual word embeddings: Often, word embeddings are induced from a source language cross-lingually (CL) into a target language. A such, Upadhyay et al. (2016) perform an empirical comparison of different approaches for inducing CL embeddings, each with a different degree of supervision: First, a simple bilingual Skip-Gram model (Luong et al., 2015) that uses word-aligned corpora to learn contexts for words in different languages; Second, a bilingual compositional model (Hermann and Blunsom, 2014), which finds bilingual embeddings for parallel sentences—each represented by the embedding of its constituent words—using minimized Euclidean length between two candidate sentences; Third, bilingual word vector training based on bilingual documents that upfront were randomly generated from a document-aligned corpora (Vulić and Moens, 2015).

Shi et al. (2015) study matrix co-factorization to learn word embeddings language-independently from distributed meaning. They first induce contexts based on word frequencies from parallel sentences. Then, they maximize similarity of word pairs in multiple languages using probabilistic machine-translation. Results in document classification show that the technique is efficient to encode CL knowledge to create CL word embeddings.

Klementiev et al. (2012) start from an annotated, well-resourced language to study word representations for joint languages. They treat word representation learning as a multitask problem where each task represents a word. Task relatedness is derived from co-occurrence statistics in bi-texts. Their approach partly outperforms MT baselines.

Cross-lingual embeddings can be understood as instance-based transfer since merging data sources from two languages modifies the distribution of words in the new embeddings space. However, when applying it to a classification problem, such as metaphor prediction, it also is an example for feature-based transfer, because we attempt to reduce the cap between the marginal contribution of the words in the embeddings representation following the conditional distributions of the labels.

Using pre-trained models in NLP tasks: Durrani et al. (2021) investigate how fine-tuning of neural models affects the learned knowledge in linguistic downstream tasks. Performing their test on pre-trained models such as BERT and RoBERTa, they use diagnostic classifiers on the layer-level and neuron-level. The authors find out that while linguistic knowledge is distributed in the entire pre-trained network, after fine-tuning it becomes localized in shallower layers, whereas deeper layers are reserved for task specific knowledge. Ahmad et al. (2021) show that explicitly providing language syntax and training mBERT using an auxiliary objective to encode the universal dependency tree structure helps cross-lingual transfer. The authors perform experiments on text classification, QA, NER, and task-oriented semantic parsing. The experiment results show that syntax-augmented mBERT boosts transfer performance with 3.9 and 3.1 points in PAWS-X and MLQA benchmarks.

Typically, TL using transformers is applied together with a fine-tuning on data samples in the target language. Hence, it is a candidate for instance-based transfer learning where the marginal distribution of the source language's instances is reweighted towards the target language.

Transfer learning in neural machine translation: Neural machine translation often approaches TL by first training a "parent" model for a high-resource language pair and then fine-tune it on a low-resource language pair ("child") by simply replacing the training corpus (Kocmi and Bojar, 2018; Zoph et al., 2016). Kocmi and Bojar (2018) find that this child model can perform better than a low-resource trained baseline even for languages with different alphabets. Similarly, Zoph et al. (2016) improve baseline models by 5.6% of BLEU score on low-resource language pairs. In a different setup, Nguyen and Chiang (2017) use parallel data from two related low-resource language pairs. A model is trained on the first language pair, then its parameters are transferred to another model where training is continued. Imankulova et al. (2019) improve TL in a Japanese–Russian pair by more than 3.7 BLEU points over a baseline. English serves as pivot language to train a multilingual model. They then fine-tune on in-domain data. Another translation example is text-to-speech (TTS) generation. To apply TTS for low-resource target languages Tu et al. (2019) transfer knowledge from a high-resource language by mapping linguistic symbols between source and target. This mapping preserves pronunciation information in the transferring process. Experiments show that 15 minutes of paired data is sufficient to build a TTS system.

In this paper, we attempt to classify German language metaphor from training a classifier using English language training data. The English language training data is represented by bilingual embeddings. In future work, we will then also test pre-trained transformers as well as techniques from machine translation.

3 Method Overview

In the following sections (i.e., Sec. 4 to Sec. 7), we present a procedure to TL for metaphor prediction. We start with a description of the metaphor corpus that we use as Gold data. Once it is completely translated and annotated for metaphor source words in the German translation (Sec. 4), we can use it for other evaluation setups too. Right now, we have 500 samples finished and use them for neighborhood retrieval (c.f., Sec. 6) and classification testing (c.f., Sec. 7) In Section 5, we introduce the source data that we build our bilingual embeddings upon and describe a merging strategy of the parallel data. We also present different approaches to handle compound metaphor sources in the target language and how they affect the distance to the English language counterparts (in the embeddings space). The latter is performed on 500 samples of the metaphor corpus. In Sec. 6, we discuss the training of the bilingual embeddings after we perform a retrieval of a metaphoric German language

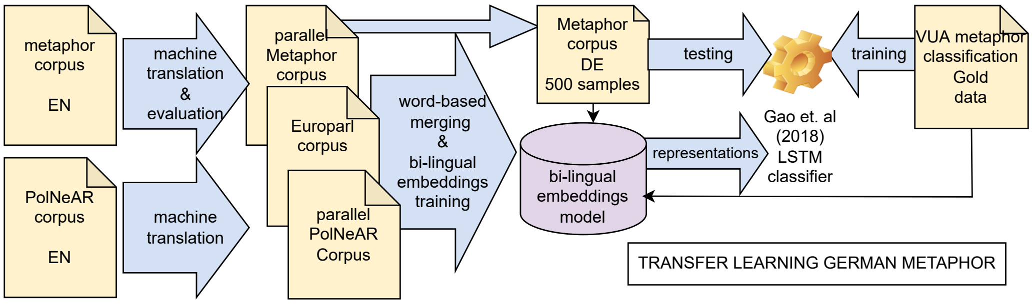


Figure 1: Overview of corpus translation and alignment in order to obtain bilingual embeddings for metaphor prediction training

word within the English language word's embeddings space. In Sec. 7 we present first results on predicting metaphor in a target language when only labeled training data in the source language is available. We use 500 already annotated samples in the target language for testing. Figure 1 shows the overview of the procedure.

4 Metaphor Gold corpus

The corpus: A first step is to create a Gold corpus to have a test set available for all sorts of techniques, be it supervised, unsupervised or transfer. Hence, we start from the corpus of Gordon et al. (2015). It origins from sources such as news articles, blog posts, and online forums. It consists of more than 1700 sentences using metaphoric language. The authors propose the use of conceptual schema to represent scenarios of metaphor usage. They recognize 70 source domains which again are grouped into 14 ontological categories. The corpus is manually validated and contains annotations for a metaphor's target, a metaphor's source, their associated linguistic and conceptual metaphors and the metaphors' lexical trigger. The linguistic metaphor annotation refers to terms from the sentence itself, so we can use this information to find the corresponding figurative label for a term. We prefer this corpus over the famous VUA corpus (Steen et al., 2010), especially because we are also interested in seeing the effect of having training and testing data from different domains (see Sec. 7). Further, having a German-translated and annotated version in place, we can add more diverse data sets to the community. A last reason is that the entire data set (once mirrored to German) offers a good sample size for tuning and evaluating further neural-based classifiers.

Corpus preparation: We prepare the data for our experiments as follows. First, we translate the sentences of the corpus into German making use of contemporary machine translation techniques.[1] We evaluate a sample of 500 sentences manually by one German native. Table 1 shows the results grouping them into three categories; i) high: denoting a perfectly translated sentence that preserves the figurative meaning while not affecting any rule of well-formed syntax nor leading to a "bad" metaphor (c.f. Harati et al. (2021) for criteria judgements on good metaphors); ii) mid: good translation with skipped metaphoric language (15) or a falsely translated (stop) word (31); and iii) low: sentence was not successfully translated (most often the last part of very long sentences). Considering the fact that the majority of sentences is very well translated, in some cases just one word is effected, and only very few translated sentences are ill-formed, we simply work with the entire data set. We also metaphor-annotate these 500 samples. Precisely, we identify the German language metaphor source word.[2]

5 Bi-text for cross-lingual embeddings

Motivation: We follow the idea that metaphoric words often stay robust or conceptual across languages (Stowe et al., 2021; Shutova and Teufel, 2010; Yan et al., 2010). To obtain more resources for languages other than English, we can apply the concept of transfer learning to make use of information of the semantic environment of words (also metaphoric words) to be transferred to the target language. So, starting with an annotated English

[1] Using Google Translate with settings source language English, target language German, and operation type document: `https://translate.google.com/?hl=de&sl=en&tl=de&op=docs`

[2] We plan to annotate the entire German part of that corpus for metaphor sources to fine-tune and evaluate transformer-based models with these Gold data too. When finished, and with the agreement by Gordon et al. (2015) we will also publish this corpus.

quality	example	#
high	EN: I will be out in the city today, feeling the [...] thrust of blood, the apple-red circulation of democracy, [...]	441
	DE: Ich werde heute draußen in der Stadt sein und den [...] Blutstrom spüren, den apfelroten Kreislauf der Demokratie, [...]	
mid	EN: [...] so vital to the smooth flow of taxation within the United States.	46
	DE: [...] die für den reibungslosen Ablauf der Besteuerung in den Vereinigten Staaten so wichtig ist.	
low	EN: [...] to assist the Government of Colombia protect its democracy from United States-designated foreign terrorist organizations [...]	13
	DE: [...] um die kolumb. Regierung beim Schutz ihrer Demokratie vor den USA zu unterstützen. ausgewiesene ausländische Terrororganisa [...]	
total		500

Table 1: Evaluation of a subset of machine translated Metaphor corpus; The medium example is well translated, but does not contain metaphor anymore

metaphor corpus (for example the metaphor corpus by Stowe et al. (2021) or the VUA corpus) and some parallel data, we can predict metaphor in German language text.

Parallel data: We run experiments using our Gold corpus of parallel metaphor to apply the concept of cross-lingual embeddings. As shown above, the technique is efficient for tasks in which semantic knowledge is needed across languages. However, our Gold data is mainly for testing purposes in the classification setup. To train bilingual embeddings, we also need bigger parallel data. We use following bigger corpora:

- The English/German part of Europarl Parallel Corpus (Europarl) (Koehn, 2005)[3]
- The training data share of the Political News Attribution Relations Corpus (PolNeAR) (Newell et al., 2018)[4] to conceive a news corpus which's content is more comparable to the one of the metaphor corpus itself. PolNeAR contains 17,292 sentences. We also translate this corpus using contemporary MT.

We combine the parallel metaphor corpus with the Europarl Parallel corpus and the PolNeAR corpus in different setups. We train the bilingual embeddings using Gensim's word2vec implementation.[5]

Merging procedure: Typically text sources for training bilingual embeddings are in a way aligned or merged (Vulić and Moens, 2015; Luong et al., 2015; Hermann and Blunsom, 2014). We generate bilingual merged text data designing a simple zip-like merging algorithm that takes the words of two sentences (English and German) as arguments. In case one sentence is longer than the other, the factor of this ratio is used to align multiple words from the longer sentence towards the shorter one. See Alg. 1 for details. We remove stop words[6] before applying the zip-merge algorithm to the Metaphor, the PolNeAR, and the whole Europarl corpus.

Algorithm 1: Merging of English and German sentences

Input: $E \leftarrow$ word token list of an English language sentence
Input: $G \leftarrow$ word token list of the German translation
Output: $EG \leftarrow$ merged token list
Ensure: $E \geq G$
$factor = round(|E|/|G|)$;
$j = 0$;
for i *in* $|G|$ **do**
  $EG = EG \cup G_i$; `/*i starting with 1*/`
  while $factor * i > j \geq factor * (i-1)$ **do**
    $EG = EG \cup E_j$;
    $j = j + 1$
  end
end

Handling compounds and derivatives: Handling compounds is a challenging matter. Our target language is famous for shipping with an extraordinary compositional nature especially concerning nouns. We count 68 compounds in our target language data set's metaphor sources (61 nouns and 7 verbs).

Cordeiro et al. (2016) handle English compound words by comparing the embedding of a compound with the embedding of its components' normalized sum. Their hypothesis is that if the angle between both embedding vectors is small then the compound's meaning is literal otherwise its meaning is idiomatic.

We decompose our compounds manually. Then, we retrieve three versions of them in the embeddings spaces that we compare later on with the English language counter word: i) the compound itself[7] (compound std), ii) the averaged vector of its components (components av.), and iii) the nor-

[3] https://www.statmt.org/europarl/
[4] https://github.com/networkdynamics/PolNeAR
[5] https://pypi.org/project/gensim/
[6] For German: https://stopwords.net/german-de/; for English we apply the stop word list delivered with the scikit-learn Python package
[7] This often falls out of vocab

malized sum of its components (Cordeiro et al., 2016) (components norm sum). For derivatives (verbs) we consider i) the finite verb form only (finite), and ii) the infinitive (infinite). We compare these vectors then with the word vector from the metaphor source of the English language text (see next section).

6 Training cross-lingual embeddings

Before we develop a supervised training setup with our data at hand (next section), it is important to learn about the potential contexts offered by cross-lingual embeddings. Therefore, we first test different setups of cross-lingual embeddings to retrieve the distances between a metaphoric word in an English language text and its German counterpart in the target language.

Using 500 manually annotated samples of our (parallel) metaphor corpus, we now retrieve the German counter word given an English language metaphoric word in the embeddings spaces trained from different parallel data setups:

- the metaphor corpus only (Metaphor); train vectors of length 150 with a min. frequency of 2 for 5 epochs
- the metaphor corpus and the PolNeAR corpus (Metaphor+PolNeAR); train vectors of length 150 with a min. frequency of 2 for 5 epochs
- the metaphor corpus & the first 100,000 sentences (to have a comparable data set) of Europarl corpus (Metaphor+Europarl 100K); train vectors of length 150 with a min. frequency of 2 for 5 epochs
- the metaphor corpus & Europarl corpus (Metaphor+Europarl); we train with vectors of length 300 with a min. frequency of 20 for 5 epochs, because this data set is much bigger then the previous data sets

Figure 2 shows the distribution of the German metaphoric word among the nearest neighbors of a metaphoric word from the English data. All corpora except the small metaphor corpus show an inverted bell curve meaning that most of the metaphors have their German counterpart among the 100 nearest neighbors or beyond their 10,000 nearest neighbors. The metaphor data (blue) rather show a bell distribution of the metaphoric words in the target language. However, we only added the blue curve for comparison reasons. The distributions of English and German metaphoric words in

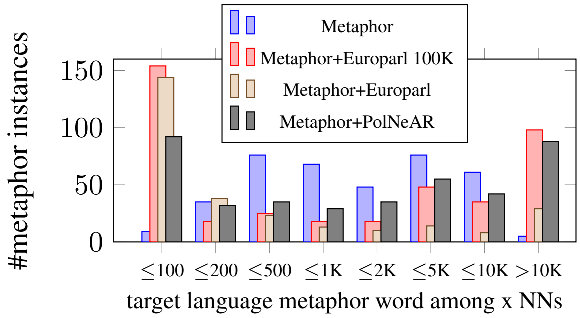


Figure 2: Distribution of metaphors which's German language metaphor source word are within the k-NNs of the English language source words

the embeddings space gives first insights into how they are represented in the bilingual embeddings' vocabulary, and hence, in the language's semantics.

Fig. 3a shows a scatter plot of the distribution of cosine similarities between the metaphors' sources in English and German respectively in the metaphor corpus. The vast majority of values is very close to one. This is especially because the vocabulary of this model is not big, and most words are in close neighborhood of the metaphor source word.

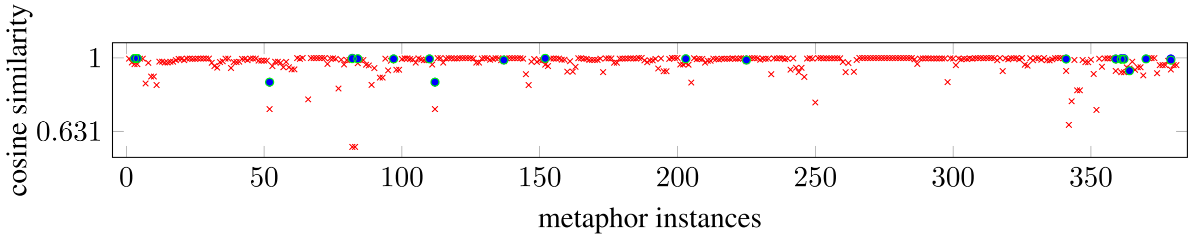
Fig. 3b shows a plot of cosine similarities distributed between the metaphors sources in English and German respectively in the metaphor corpus and PolNeAR. As PolNeAR is about ten times as big as Metaphor, we can see the data points are moving more towards zero being not as similar anymore.

Fig. 3c shows cosine similarities distributed in Metaphor and Europarl 100K. We encounter a much lower oov-rate (min-freq of 2). We also see that the normalized sum for component combination achieves higher similarities with the English metaphor sources than the averages do. Even though we do not have many data points here, we still learn that compounds are somewhat difficult to associate to the English source words. Hence, we plan to test the impact of decomposed compounds in the test data once our entire Gold corpus is finished[8]

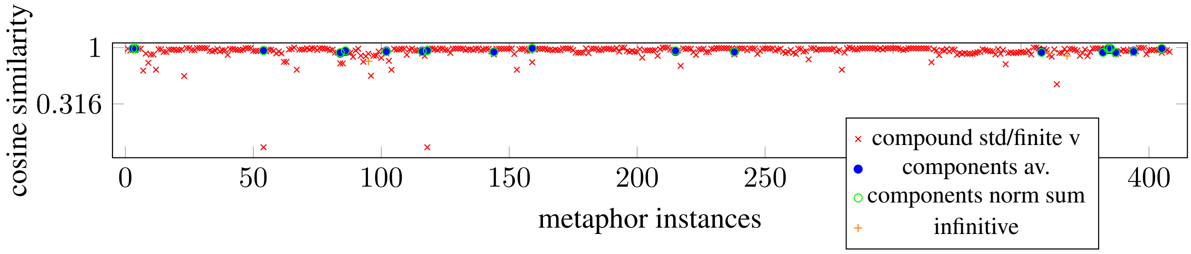
In the next section, we use our bilingually trained embeddings[9] in a TL-classification task. We find out experimentally that a mix of Metaphor, PolNeAR, and Europarl with a minimum frequency of 5 and lower-cased embeddings sources covers most

[8]For other corpus combinations we do not show scatter plots since similarity decreases with vocabulary growth.

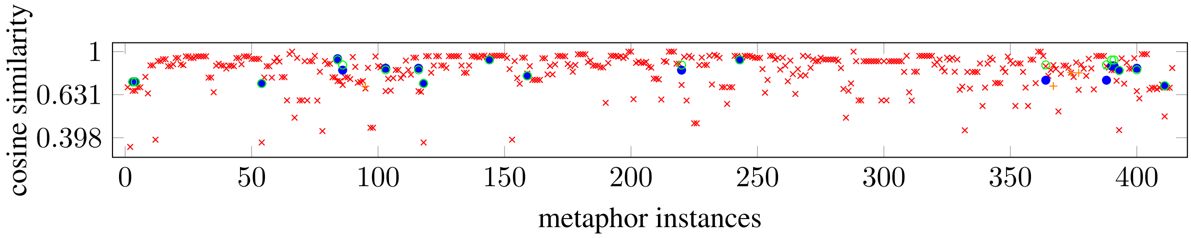
[9]Using a window of 5, five epochs, and 300 dimensions to match default values



(a) **Metaphor:** 381 data points displayed; 104 oov; 15 no metaphor in target; y-axis logarithmic



(b) **Metaphor+PolNeAR:** 408 data points displayed; 77 oov; 15 no metaphor in target language; y-axis logarithmic



(c) **Metaphor+Europarl 100K:** 414 data points displayed; 71 oov; 15 no metaphor in target; y-axis logarithmic

Figure 3: Distribution of cosine-based similarities between a metaphor source word in EN and DE

of the vocabulary in the training and test data (c.f., Tab. 2).

embedding sources	min f	#voc
Metaphor+PolNeAR	2	42,353
Meta+Europarl 100K lc	5	30,768
Meta+PolNeAR+Euro 100k	5	42,008
Meta+PolNeAR+Euro 100k lc	5	39,229
Meta+Europarl	20	68,506
Meta+PolNeAR+Euro lc	5	139,356

Table 2: Vocab sizes of different embeddings sources; using Metaphor, PolNeAR, Euro(parl) (100,000 sentences) l(ower)c(ased) next to min(imum) f(requnecy) and voc(ab size)

7 TL with cross-lingual embeddings

Experimental setup: Inspired by Gao et al. (2018), we use the VUA corpus (Steen et al., 2010) together with our bilingual embeddings to perform cross-lingual metaphor prediction. This means, we train the model from Gao et al. (2018) as presented in their work (the authors use GloVe embeddings (Pennington et al., 2014) and ELMo embeddings (Peters et al., 1802) with a bidirectional LSTM classifier), then we use our German metaphor data set as test set. Our test data consist of a balanced data set of sentences labeled with 1 (when it contains a main verb that is metaphoric; 259), and with 0 (when it is not; 198). The index of the respective verb is handed over as well.

We run four setups: i) the baseline approach training/testing on VUA using GloVe embeddings (Gao et al., 2018) (no transfer); ii) the same setup using our embeddings instead of GloVe (no transfer); iii) our embeddings testing on the English part of the Metaphor corpus (no transfer); and iv) our embeddings testing on the German part of the Metaphor corpus (transfer).

Results and discussion: Table 3 shows that our embeddings do not address the vocabulary of the training and testing data as well as GloVe does. Still, the bilingual embeddings are capable to represent contexts well as F1-scores rather increase than drop for English (row 2 and 3 compared to row 1). Accuracy, however drops drastically especially

embed	voc addr	voc size	train	sample size	test	sample size	val f1	p	r	f1	ac
GloVe (reprod.)	17,941	18,695	VUA	17,240	VUA	5,873	57	59	53	56	75
M+P+E	11,480	18,695	VUA	17,240	VUA	5,873	52	56	69	62	75
M+P+E	10,862	17,301	VUA	17,240	M-En	480 (284:196)	52	65	66	65	59
M+P+E	11,845	19,567	VUA	17,240	M-De	457 (259:198)	54	60	22	33	48

Table 3: Results (%) of TL classification in metaphor prediction using our embed(dings model): M(etaphor)+P(olNeAR)+E(uroparl) 1.9mio lower-cased; voc(ab) addr(essed); voc(ab) size

while applying the English-trained model in German (row 3 and 4). This might be the case since our testing data set is better balanced than the VUA data set. On the other hand, metaphoric contexts are not as well represented for German as they are for English in the model, especially since we did not word-align the data even though positions of verbs differ in both languages. [10] Looking into samples, we found that the especially low recall is caused by a lot of verbs not used very figuratively, such as "Waffenrechte verteidigen/schützen" (defend/protect gun rights). We plan to investigate these issues in detail to refine our choices and methods for training bilingual embeddings.

We did not use pre-trained bilingual embeddings even though existing work often comes with links to data and code (c.f, Luong et al. (2015); Hermann and Blunsom (2014)). However, these data often is difficult to collect as links are not available, broken or regeneration is laborious. Further, often bilingual embeddings are trained on Europarl which is not necessarily the domain, we can hope to find a lot of metaphoric language—also a problem in our approach as we use Europarl data too. During our work we also learned that adding source data from the news domain to our embeddings data reduces distances in the embeddings space (c.f., Fig. 3b).

8 Next steps

A next step is to predict metaphoric language in a target language using pre-trained transformer models and our Gold data for fine-tuning for example in a classification task. For this task, the embeddings representation of a sentence and the metaphor's source word is given, and the metaphoric word of the target language needs to be predicted.

Another step might be applying TL methods of neural machine translation (e.g., Kocmi and Bojar (2018)). As we learned in Sec. 2.2 usually a neural model is trained on a high-resource language pair and tested on a low-resource pair. In our setup, we could encounter this using a sequence labeling model trained on an English language metaphor corpus and combine it with a (pre-trained) translation model of English and German. As sources for the translation and evaluation part, we might also consider to use the parallel data from Common Crawl EMNLP (2018). Figure 4 shows an overview on the technique considering language model and tagging model probabilities as common translation setups do.

[10] We also tested our approach using bilingual embeddings from upfront word-aligned (Jalili Sabet et al., 2020) data. However, test F1-score remains below 10%. We belief that the n:m relations of words make it difficult for the classifier to identify the semantic in the target language well enough.

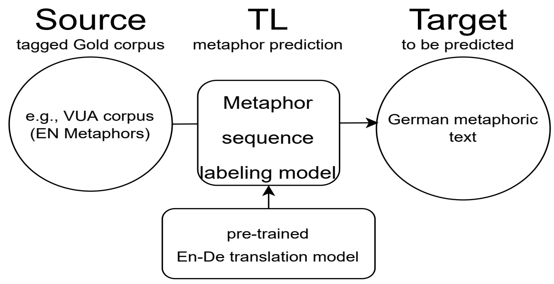


Figure 4: Overview on metaphor detection using a sequence labeling transfer learning technique

9 Conclusion

In this paper, we presented an overview of transfer learning techniques structured in a twofold manner: i) types of transfer learning, and ii) transfer learning techniques from a task-oriented perspective. We presented first steps towards the application of modern transfer learning techniques towards metaphor prediction in German language text. The experiments make clear that successfully training bilingual embeddings depends on the vocabulary coverage of the source texts. We furthermore are in the process of annotating a parallel corpus (EN-DE) of metaphor starting from a pre-existing English language corpus, which we plan to use as a Gold data set to test transformer-based models.

10 Limitations

Our first results show very low performance considering a guessing baseline of about 50%. We think this is mainly caused by the limited embeddings data we have available. Also, the lack of word alignments might cause difficulties. However, as demonstrated, the task is very complex given all the constrains that need to be fulfilled upfront (such as Gold data set, suitable TL-technique, bilingual resources). We consider to look further for parallel data sources and develop strategies to generate parallel sources, e.g., by back-translation (Dhar et al., 2022) before we go ahead applying other TL-learning techniques. We also need to establish a way to incorporate findings on compound and infrequent words into the creation of the embeddings representation. However, we did not do this yet, because we had to manipulate the primary data for this purpose.

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.

Wasi Uddin Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual bert for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4538–4554. ACL.

Azin Asgarian. 2018. An introduction to transfer learning. Georgian Impact Blog, online. Accessed: Oct. 2022.

Azin Asgarian, Parinaz Sobhani, Ji Chao Zhang, Madalin Mihailescu, Ariel Sibilia, Ahmed Bilal Ashraf, and Babak Taati. 2018. A hybrid instance-based transfer learning method. *arXiv preprint arXiv:1812.01063*.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997.

Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2008. Translated learning: Transfer learning across different feature spaces. *Advances in neural information processing systems*, 21:353–360.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2022. Evaluating pre-training objectives for low-resource translation into morphologically rich languages. In *The 13th Conference on Language Resources and Evaluation*, pages 4933–4943. European Language Resources Association (ELRA).

Lettie Dorst. 2015. More or different metaphors in fiction? a quantitative cross-register comparison. *Language and Literature*, 24:3–22.

Lixin Duan, Dong Xu, and Ivor Tsang. 2012. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660*.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep nlp models? *arXiv preprint arXiv:2105.15179*.

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.

Parastoo Harati, Chris Westbury, and Milad Kiaee. 2021. Evaluating the predication model of metaphor comprehension: Using word2vec to model best/worst quality judgments of 622 novel metaphors. *Behavior Research Methods*, 53(5):2214–2225.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 58–68. ACL.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643. ACL.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271.

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

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

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

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

Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip. 2013. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

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.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210.

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. 1802. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 567–572. ACL.

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)*.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

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.

Tao Tu, Yuan-Jui Chen, Cheng-chieh Yeh, and Hung-Yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1661–1670. ACL.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372. ACM.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Ding Yan, Dirk Noël, and Hans-Georg Wolf. 2010. Patterns in metaphor translation: a corpus-based case study of the translation of fear metaphors between english and chinese. *Using Corpora in Contrastive and Translation Studies. Newcastle: Cambridge Scholars Publishing*, pages 40–61.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.