Proceedings of SIGUL2022 @LREC2022, pages 176–184
Marseille, 24-25 June 2022

Investigating the Quality of Static Anchor Embeddings from Transformers for Under-Resourced Languages

Pranaydeep Singh, Orphée De Clercq, Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{firstname.lastname}@ugent.be

Abstract

This paper reports on experiments for cross-lingual transfer using the anchor-based approach of Schuster et al. (2019) for English and a low-resourced language, namely Hindi. For the sake of comparison, we also evaluate the approach on three very different higher-resourced languages, viz. Dutch, Russian and Chinese. Initially designed for ELMo embeddings, we analyze the approach for the more recent BERT family of transformers for a variety of tasks, both mono and cross-lingual. The results largely prove that like most other cross-lingual transfer approaches, the static anchor approach is underwhelming for the low-resource language, while performing adequately for the higher resourced ones. We attempt to provide insights into both the quality of the anchors, and the performance for low-shot cross-lingual transfer to better understand this performance gap. We make the extracted anchors and the modified train and test sets available for future research at https://github.com/pranaydeeps/Vyaapak

Keywords: cross-lingual transfer, bilingual lexicon induction, natural language inference

1. Introduction

Despite the great progress witnessed in recent years for various NLP tasks, low(er)-resourced languages are often lagging behind because of data scarcity. To overcome this lack of resources, researchers have started to investigate the use of cross-lingual information, where knowledge or data from a rich-resourced language, like English, is used to improve the modeling in a low(er)-resourced target language. With the new dawn of extremely data hungry (pre-trained) transformers, the field of cross-lingual knowledge transfer has become even more effective, since large pre-trained models are not always available for a certain language or task.

The idea of cross-lingual embeddings originally stems from the idea of Mikolov et al. (2013) that vector spaces in different languages share a certain similarity, and that a projection can be learned from one language to another. A lot of research has been proposed to perform cross-lingual alignment (see Section 2 for an overview). The most recent approaches incorporating contextual embeddings, such as multilingual BERT (mBERT, Devlin et al. (2019)) and XLM (Conneau and Lample, 2019) apply joint training on multiple languages, obtaining very promissing results for a wide range of cross-lingual tasks. Main drawback of these approaches is that they require a huge amount of processing time and power, which makes them almost impossible to retrain for additional languages. In addition, research has shown that low-resourced languages are under-represented in joint models like mBERT and perform poorly on downstream tasks compared to high-resourced languages (Wu and Dredze, 2020).

The approach under investigation here has initially been proposed by Schuster et al. (2019). They demonstrate that contextual embeddings can be treated as having a static anchor component, and a dynamic context component for every token. In this paper, we revisit and investigate the potential of this static anchor component for the cross-lingual transfer of transformer representations for under-represented languages, Hindi in this case. We compare all results with a set of control target languages having more resources and which are either closely (Dutch) or more distantly related (Russian, Chinese) to the source language English. Although a language like Hindi has a large number of native speakers (around 370 million worldwide), NLP researchers consider a language to be low-resourced when it is difficult to gather corpora or tools for that specific language (e.g. the size of the Wikipedia available for training language models (Wu and Dredze, 2020)).

We extend the original anchor-based approach in several ways. First, up to date the original approach has not been evaluated for BERT or other language models from the transformer family since it was proposed in a pre-transformers era. Second, it has only been evaluated on a set of higher-resourced Western European languages, and not on under-resourced languages, such as Hindi. Third, the original work demonstrated its use case solely for dependency parsing, while we evaluate the quality of the anchors for two sets of tasks: (1) monolingual tasks: Word Polarity Prediction, and (2) cross-lingual tasks: Bilingual Lexicon Induction (a lexical task) and zero-shot Natural Language Inference (a sentence-based task). For each task, we compare our approach to the state-of-the-art methodologies. We provide a detailed overview of all experimental results and also attempt to analyze in detail the inherent drawbacks and failures of the approach.

The remainder of this paper is organized as follows. Section 2 describes the related research on cross-lingual approaches, whereas Section 3 further elaborates the anchor-based approach we extended to obtain cross-lingual representations from pre-trained transformers. Section 4 gives an overview of the experimental setup and results, both for the mono and cross-lingual downstream tasks. Section 5 provides a qualitative analysis and discussion, while Section 6 ends this paper with concluding remarks and indications for future research.

2. Related Research

There are various research strands using cross-lingual information to circumvent the lack of resources in a given target language.

A first line of research uses machine translation (MT) systems to map lexicons or labeled data to other languages (e.g., (Mihalcea et al., 2007) for the task of Sentiment Analysis). Balahur and Turchi (2014), however, showed that working with translated data implies an incremented number of features, sparseness and noise in the data for classification. They also revealed that the quality of these methods largely depends on the availability of large parallel corpora for training the MT system, which are often lacking for low-resourced languages. Related approaches only use parallel data without building machine translation systems. Rasooli et al. (2018) used annotation projection to project supervised labels from the source languages to the target language and a direct transfer approach to develop sentiment analysis systems.

Other approaches extract paired sentences from large parallel corpora to learn bilingual embeddings. Chandar et al. (2014), for instance, explored the use of autoencoder-based methods for learning vectorial word representations that are aligned between the two languages without relying on word-level alignments. They reported state-of-the-art performance for the task of cross-language text classification. In sum, all these approaches require large amounts of high-quality parallel data, which are often lacking for low-resourced languages.

Another promising line of research, one that does not require large parallel corpora, are cross-lingual embeddings. These cross-lingual embeddings, which are obtained by mapping monolingual word embeddings into a common space, have already been successfully applied for low-resourced languages (Duong et al., 2016). The concept entails the possibility of learning a perfect mapping by traversing between vector spaces in different languages. In other words, by creating monolingual spaces and then learning a projection from one language to another the need for large parallel corpora for cross-lingual supervision can be eliminated. Mikolov et al. (2013) attempted to learn a linear mapping from one space to another and optimized the performance by using the most common words from both languages and by using a bilingual lexicon to guide the learning of the mapping in the right direction. As large bilingual lexicons are often not available for low-resourced languages or specific domains, there was a need to either completely eliminate or drastically reduce the size of the required bilingual lexicon. Artetxe et al. (2017) further explored these ideas by using a combination of back translation and denoising. This approach was, however, severely lacking in terms of performance as compared to a method with cross-lingual signals. The advent of adversarial networks brought on some unique ideas which opened up a lot of new research directions: a discriminator is trained to identify whether an embedding originates from a source language or a target language and a mapping is trained to fool the discriminator. The underlying principle is that there is an orthogonal matrix W, which can transform embeddings in one language to embeddings in another language.

With the arrival of the new generation of language models, contextual embeddings came into the picture. Contextual embeddings significantly enhanced word and sentence representations, and improved upon previous methods of cross-lingual alignment like MUSE (Lample and Conneau, 2019) and VecMap (Artetxe et al., 2018) due to their dynamic nature. Multilingual BERT (mBERT, Devlin et al. (2019)) and XLM (Conneau and Lample, 2019) were jointly trained for Masked Language Modelling on 104 languages and significantly outperformed previous approaches for a variety of zero-shot cross-lingual tasks. While joint training is an excellent solution, it is computationally expensive to train and not receptive to new languages after the initial training. A number of recent works (Wu and Dredze, 2020; Wang et al., 2020) investigating mBERT have also uncovered that under-resourced languages have much poorer representations compared to the higher-resourced languages, making these models not the optimal choice when working with a low-resource language.

Artexte et al. (2020) introduce another clever alternative to joint training (mBERT, XLM), by freezing the encoder layers of a transformer after the initial pre-training, and re-learning only the embeddings on a target language. This results in a very similar performance to mBERT while keeping the training time significantly lower. Schuster et al. (2019), for their part, treat contextual embeddings as having a static anchor component, and a dynamic context component for every token. This once again enabled the static components to be aligned with methods like MUSE. Tran et al. (2020) proposed a further improvement on the joint training direction of research, by forcing foreign language embeddings to be initialized in the same space as the source language, thus increasing the performance of mBERT and XLM.

In this paper, we seek to investigate viable approaches to zero-shot cross-lingual transfer of transformer representations for a lower-resourced and often under-performing language, namely Hindi. At the same time we wish to compare the performance of these approaches on higher-resourced languages from different families. To this end, we revisit the anchor-based approach of Schuster et al. (2019) which decomposes contextual embeddings into anchors and contexts. Given that this original approach has only been validated on ELMo (Peters et al., 2018), we investigate the scalability of this method on modern transformers such as BERT and RoBERTa (Liu et al., 2019). In order to assess the viability of this approach on Hindi in different settings, we perform detailed experiments for three different downstream tasks.

3. Static Anchors from Transformers

Even though approaches like RAMEN (Tran, 2020) and MonoTrans (Artetxe et al., 2020) have replaced the older, orthogonal alignment with Procrustes refinement strategies, these newer approaches are solely designed for certain architectures requiring additional training steps. In this paper we choose to investigate an approach which is intuitively sound and model-agnostic. The approach in question, henceforth referred to as Cross-lingual ELMo (Schuster et al., 2019), theorizes that the average for all contextual embeddings of a word over a large corpus adequately represents a static anchor for the token in question. Given a source language s and a target language t, the objective of the classical alignment methods is to learn a transformation

$$E_{s,t} \approx W^{s \to t} E_{s,s} \tag{1}$$

where $E_{s,s}$ represents the embeddings of the source language in their original space, while $E_{s,t}$ represents the embeddings of the source language in the target language's multi-dimensional space. For classical word embeddings like word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2016), this becomes a simple optimisation problem for an orthogonal matrix W. VecMap achieves this by maximizing for similarity over a sparse seed dictionary (which can be initialized with zero supervision or using identical words if a seed dictionary is not available), and iteratively improving the dictionary and relearning the alignment after each optimisation step. MUSE achieves the same objective by initializing W using an adversarial objective, where W is optimized such that a discriminator model is unable to differentiate between the embeddings originating from $E_{t,t}$ and $WE_{s,s}$.

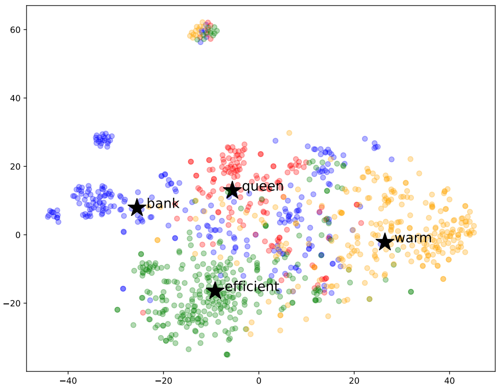


Figure 1: Distribution of token embeddings from all Wikipedia contexts for the words *bank, efficient, queen* and *warm*, and their respective static anchors (⋆).

However, the dynamic nature of the embedding spaces E in the case of transformers makes the solutions slightly more complicated and requires some assumptions to simplify the problem. To obtain an approximation of the embedding spaces $E_{s,s}$ and $E_{t,t}$, for a token i in the context c,

$$e_{i,c} = A_i + e_{\hat{i,c}} \tag{2}$$

where A_i is the fixed Anchor for the token i obtained by averaging embeddings over all available contexts c, while $e_{\hat{i,c}}$ is the additional context component of the embedding. This decomposition means that the complete embedding space $E_{s,s}$ once again can be simplified as a static space $A_{s,s}$, the space of all anchors for a source language s. The outcome of the anchor extraction approach is shown in Figure 1 for four example words (*bank, efficient, queen* and *warm*). The individual dots represent the embeddings of the tokens in various contexts from the Wikipedia corpus, while the ⋆ represents their obtained anchors. In their paper Schuster et al. (2019) demonstrated that for the ELMo embeddings the point clouds for individual tokens can be seperated much more distinctly and thus may result in better anchors. However, if we look at Figure 1, more intersecting clouds can be observed for our BERT embeddings.

After the static anchor space is obtained, a transformation

$$A_{s,t} \approx U^{s \to t} A_{s,s} \tag{3}$$

can then be learned with methods like MUSE and VecMap, to align monolingual anchors with their counterparts in other languages. Figure 2 illustrates the outcome of this alignment for the same four words in English and Dutch: we indeed observe that the anchor in English (⋆) is well-aligned with the anchor in Dutch (△). However, 'bank' being a homonym in English interferes with the alignment of its different meanings in Dutch. This again in contrast to the ELMo anchors where homonyms were often found to be resolved succesfully.

While this alignment method for dynamic contextual embeddings has been shown to perform well

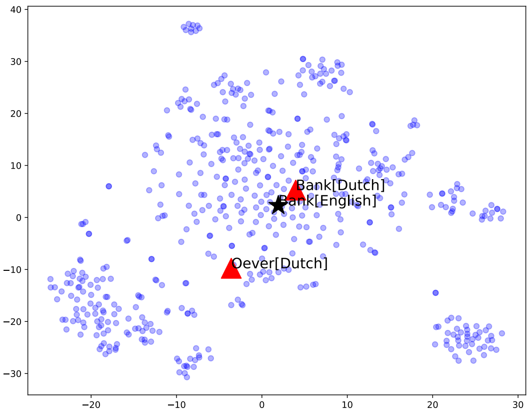


Figure 2: Homonyms: different meanings of the word ‘bank’ in Dutch (financial: *bank* /vs/ river bank: *oever*) are anchored similarly to ‘bank’ in English.

using ELMo anchors for dependency parsing, we further probe the potential of this methodology for transformer-based architectures to under-resourced languages. Below, we perform detailed experiments to probe the quality of the anchors, first in a monolingual setting to judge the quality of their pre-alignement, then in a cross-lingual setting by aligning anchors with VecMap and testing them for the tasks of Bilingual Lexicon Induction and and Zero-shot Natural Language Inference.

4. Experimental Setup and Results

The initial step for both sets of experiments is identical, i.e. the extraction of anchors from a BERT-based model. We aim to study the anchors for a wide variety of BERT-based transformers. While for English[1], Hindi[2] and Chinese[3], anchors extracted from more standard BERT models, we relied on RuBERT (Kuratov and Arkhipov, 2019) for Russian, which is a cased BERT model initialized with mBERT, and on Robbert (Delobelle et al., 2020) for Dutch, which is a RoBERTa-based architecture. We use these pre-trained LMs, along with a random subset (1 million sentences) of Wikipedia in the respective languages, to extract embeddings for the 50,000 most common words in the corpus. All the different contexts are then averaged to obtain the anchors as described in Section 2. We perform all described experiments on a singular Tesla V100 (16GB) which takes about 30 hrs per language. Since this is the only major bottleneck in the experiments, we make the obtained anchors available for use.

4.1. Monolingual Evaluation

To judge the quality of the anchors’ pre-alignment, we perform baseline experiments to compare them with FastText embeddings trained on an identical Wikipedia corpus. We train both sets of embeddings with an additional linear layer for classificitation, viz. to predict the polarity of words contained by the Multilingual Sentiment Lexicon (Chen and Skiena, 2014). We use 2,000 random words from the lexicon for training and 400 for testing for each language (except for Chinese (*) where we only had 1000 words for training). The experiments are performed for all 5 languages used in the cross-lingual setup, English (EN), Hindi (HI), Dutch (NL), Russian (RU) and Chinese (ZH). Working with a token-based polarity prediction instead of sentence-based sentiment analysis made more sense for this evaluation since we aim to study the lexical strength of the embeddings before proceeding to more complicated tasks.

The scores for the monolingual setup are shown in Table 1. There is a significant performance gap between FastText and the obtained anchors for most languages except for Russian and Chinese, with Chinese being the only language where the static anchor approach outperforms FastText. The performance for the anchors was found to be especially poor for Hindi and Dutch, while the FastText counterparts remain more or less consistent for all languages. The results clearly demonstrate that on a purely lexical basis, FastText embeddings are still quite superior, even for an under-resourced language like Hindi.

Language	FastText	Static Anchors
EN	**0.8425**	0.7575
HI	**0.8125**	0.5625
NL	**0.7300**	0.5750
RU	**0.7575**	0.7175
ZH*	0.5200	**0.5780**

Table 1: Results for the Monolingual Setup (word polarity predictions) for the five considered languages: English (EN), Hindi (HI), Dutch (NL), Russian (RU) and Chinese (ZH)

4.2. Cross-lingual Evaluation

4.2.1. Bilingual Lexicon Induction

For the first part of the cross-lingual evaluation, we perform Bilingual Lexicon Induction (BLI) experiments for four language pairs, for each pair using English as both a source (EN-XX) and target language (XX-EN). All datasets have been derived from the MUSE bilingual dictionaries[4]. Since our intention is to evaluate contextual models, the respective MUSE train and test sets had to be reduced to accommodate for the smaller BERT sub-word based vocabularies as compared to the

[1] https://huggingface.co/bert-base-uncased

[2] https://huggingface.co/monsoon-nlp/hindi-bert

[3] https://huggingface.co/bert-base-chinese

[4] https://github.com/facebookresearch/MUSE

	EN-HI	HI-EN	EN-NL	NL-EN	EN-RU	RU-EN	EN-ZH	ZH-EN
FASTTEXT EMBEDDINGS WITH VECMAP								
Full Train Set	0.5679	0.7098	0.8604	0.8467	0.6465	0.8137	0.8325	0.549
1k Supervision	0.4864	0.5268	0.8234	0.7660	0.5525	0.7561	-	-
ALIGNED ANCHORS WITH VECMAP								
Full Train Set	0.4955	0.5994	0.6382	0.7350	0.6210	0.8043	0.8010	0.4510
1k Supervision	0.3620	0.2997	0.2300	0.3860	0.3276	0.5940	-	-

Table 2: BLI Results for the four language pairs, including English both as source and target language.

Model	**HI**	**RU**	**ZH**
XNLI Transfer Learning Baseline	0.563	0.578	0.588
mBERT (Devlin et al., 2019)	0.600	0.638	-
XLM (MLM) (Lample and Conneau, 2019)	0.657	0.731	0.719
MonoTrans (Artetxe et al., 2020)	**0.660**	0.704	0.703
RAMEN (Tran, 2020)	0.656	**0.736**	**0.737**
CL ELMo (Schuster et al., 2019)	0.548	-	-
CL-anchor-BERT	0.583	0.644	0.662

Table 3: Results on the Zero-Shot XNLI Test Set

FastText or word2vec variants. Using the full dictionaries would be misleading, since, for example, for Russian, our model was only able to use around 3500 samples for training, as compared to the 5000 available in the full train set. To keep the comparisons consistent, we evaluated the two methods incorporating static FastText embeddings (VecMap and MUSE) on the reduced train/test sets as well, and make the reduced dictionaries available[5] for reproducibility. Two sets of experiments have been performed for each language pair: one with the full train set, and a second one where only 1000 samples are available for supervision, (except for Chinese where the full train set consisted of less than 1000 entries, so a run with 1000 samples was not possible). We use FastText vectors aligned with the same hyperparameters as the anchors, using VecMap for comparison.

Table 2 lists the accuracy scores for the BLI experiments. The anchor alignment methods again fail to compete in lexical strength with the SOTA VecMap alignments using FastText, except for Russian where the two methods perform quite similarly. A reason why FastText embeddings align significantly better could be attributed to the isomorphism assumption. Vulić et al. (2020) pointed out that two sets of embeddings are more likely to be isomorphic given similar environmental factors, like similar amounts of training data, time and parameters. This makes FastText very robust since embeddings for all the languages are trained in a near identical fashion.

4.2.2. Zero-Shot Natural Language Inference

In our final evaluation, we use the aligned anchors in a basic setup for zero-shot cross-lingual NLI using the XNLI (Conneau et al., 2018) dataset. As this dataset does not include Dutch, we perform the experiments for Hindi, Russian and Chinese. We first fine-tune a classifier using the English train set, with the language model fully frozen to prevent the embeddings from being altered, since the alignment matrix $W^{EN \to TRG}$ was obtained for the embedding space prior to the training step. In a second step, we use the embeddings for a transformer from the target language, using the alignment matrix to transfer the embeddings to the shared space, and use the pre-trained classifier to perform zero-shot NLI in the target language. We use a learning rate of $1e-5$, gradient accumulation for every 2 steps for a batch size of 8, and train for a total of 5 epochs for the English training phase.

We report results for the anchor-based systems, CL-anchor-BERT, for all languages, as well as results for other state-of-the-art cross-lingual methods in Table 3. We were unable to find ELMo models for Russian and Chinese, which is why these scores are only reported for Hindi. The results reported for MonoTrans, XLM and RAMEN are of the variants of the models that use no parallel corpus since the approach investigated in this paper also does not require a parallel corpus.

As can be seen in the results, CL-anchor-BERT outperforms the XNLI transfer learning baseline for all languages in question, but fails to close the gap on the state-of-the-art approaches XLM (Joint training SOTA approach), and MonoTrans/RAMEN (cross-lingual transfer SOTA approach). It is a key detail that all of the listed SOTA approaches do fine-tune the language model for the English pre-training step, while the anchors approach works with a frozen encoder, which potentially explains the gap in performance. Another potential cause for this can be the dynamic context of the embeddings being impactful for methods like RAMEN and MonoTrans, whereas CL ELMo, and by extension CL-anchor-BERT, only

[5] https://github.com/pranaydeeps/Vyaapak

use the static anchors to learn the alignment matrices, which could be a hindrance when used with context-rich BERT embeddings. It is also worth noting that CL-anchor-BERT significantly outperforms the previously used CL ELMo variant, hence also proving that the static anchor hypothesis does indeed extend to BERT and outperforms results on ELMo for Hindi.

5. Discussion

Based on the results a few observations can be made. Firstly, for the BLI evaluation, we note that with the anchor-based approach, the transfer from English is significantly harder than just relying on English as the target language, especially for Hindi and Russian. Another outcome is that the drop in performance for the 1,000 training samples experiments seems to be consistently higher for the anchor alignments compared to FastText. This could be attributed to the larger vocabulary of FastText allowing the alignment refinement steps to have a better understanding of the embedding space, thus making the anchor-based approach only viable with slightly larger seed dictionaries. This can obviously be mitigated by expanding the vocabulary of the anchors, but will exponentially increase the compute bottleneck for anchor extraction.

In order to gain more insights into the ouput of our approach, native linguists performed a qualitative error analysis on the BLI output of the first 100 instances of the test sets of Hindi, Dutch and Russian. Interestingly, we found that even though these three languages are far apart, they exhibit similar errors. Figure 4 represents the distribution of the error categories per language. As can be observed, the largest error category in Hindi constitutes nonsensical words, a problem likely caused again due to the BERT sub-word tokenization not being perfectly suited for under-represented languages. For Russian, especially morphological and syntax-related errors prevail (the latter has mostly to do with different cases or inflections of nouns, a typical difficulty of the Russian language). The other error types are related to semantics (antonyms, synonyms, polysemous words). An important category (especially in Hindi and Russian) are words that are no real translations, but are semantically related (example EN-HI: ‘chicken’ was translated to *elephant*, example EN-RU: ‘promise’ was translated to *hope*, example EN-NL: ‘inches’ was translated by *meters*, which is actually the Dutch standard distance metric).

In Figure 3 we, also attempt to visualize some selected embeddings that have been correctly (green) and incorrectly (red) aligned for Hindi, Dutch and Russian using PCA dimensionality reduction. The embeddings in blue are the source words. The visualizations demonstrate (again) that a lot of the mistakes can be attributed to semantics, as well as ambiguity in the test set (e.g. ‘bladen’ in Dutch can be interpreted as both ‘sheets’ (*of paper*) and ‘leaves’ (*of a tree*), but only ‘sheets’ is accepted by the gold standard test set). During the qualitative error analysis lots of such translations were indicated as missing from the gold standard.

Secondly, for the XNLI evaluation, we performed an analysis of the mistakes made by the CL-anchor-BERT model where MonoTrans and RAMEN were often found to be correct. We observed that most of these errors occurred for sentences containing words with less than 10 samples in the validation set of Hindi Wikipedia that was use for the anchor extraction phase. This means these instances resulted in unrefined anchors and therefore, by extension, poor alignments. This issue also potentially correlates with the frequent semantically rooted mistakes found in the BLI evaluation (such as *Persia* was was translated as Iran in Hindi). This problem could be solved by adding more monolingual data (from Common Crawl, for example) for the anchor extraction step. We also noticed that for cases where the anchors are sufficiently refined – with more than 50 occurrences of the token – CL-anchor-BERT is more consistent than MonoTrans and RAMEN. Figure 5 shows example sentences from the test set, with words occurring less than 10 times marked in red. As can be expected, the marked words have poor anchors, thus compromising the sentence representations. A manual analysis of a random sample of 20 test sentences containing no tokens with less than 50 occurrences showed that CL-anchor-BERT correctly predicts 16 instances, while MonoTrans and RAMEN correctly predict 13 and 12 instances, respectively. This demonstrates that the anchor extraction and alignment methodology has the same potential as any other proposed approach to convert a transformer from one language to another, provided that enough data is available to extract high-quality anchors.

Our final point of discussion attempts to justify the lower performance for Hindi (and by extension other under-resourced languages). In the past a possible explanation for this has been that the sub-word tokenization scheme does not benefit languages like Hindi and Urdu, which has already been studied extensively by Wu and Dredze (2020) and Wang et al. (2020). Moreover, reference can also be made to the limits of relying on unlabelled monolingual data. Since most methodologies use the Wikipedia and/or the Common Crawl corpus as initial pre-training data, the performance of under-resourced languages can be justified by directly comparing their performance as a function of the amount of available monolingual data. To this end, we compared the test accuracy of a language for the XNLI dataset using the MonoTrans methodology, with the number of pages available in the language’s Wikipedia. Figure 6 shows a significant correlation (R^2 value of 0.882 for the trendline) between the availability of monolingual data and the XNLI test accuracy for the MonoTrans SOTA methodology (in %). It is interesting to note that two languages as varied

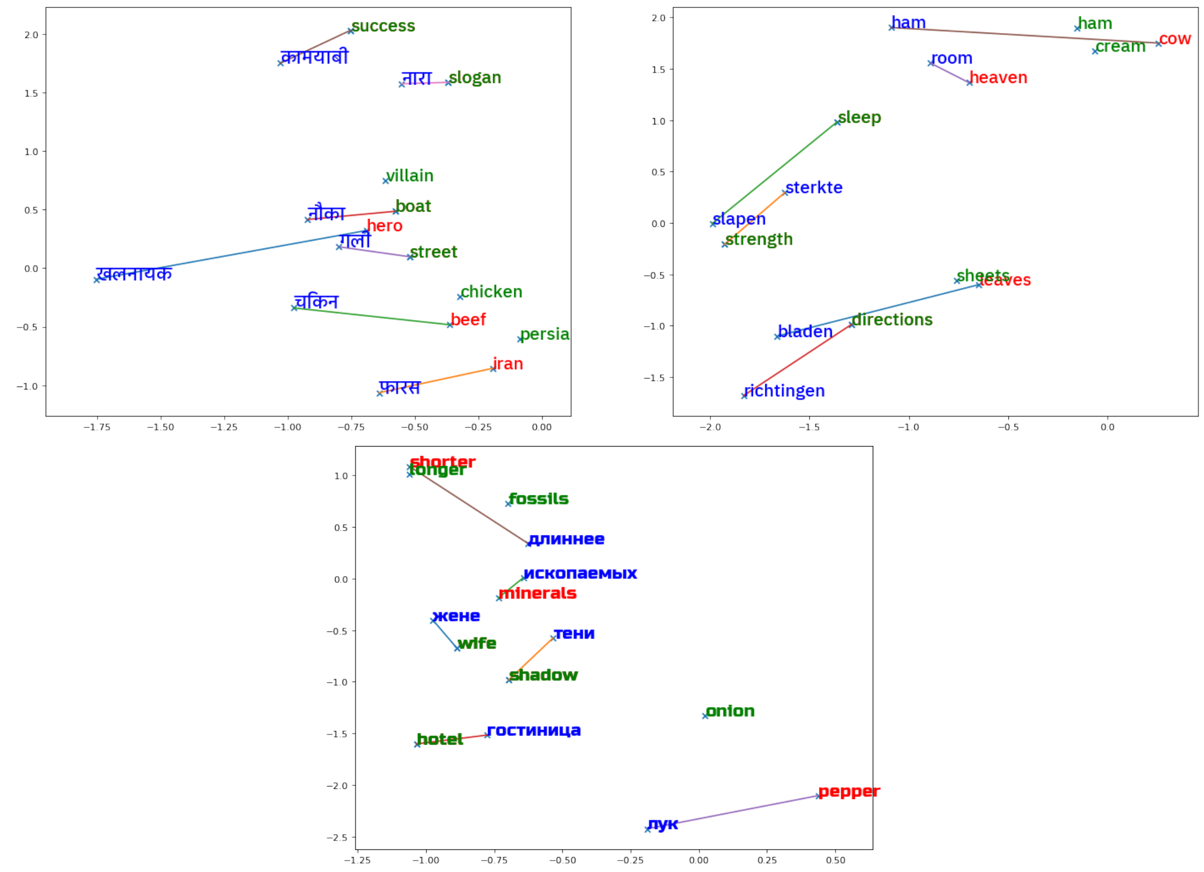


Figure 3: Illustration of the Hindi, Dutch and Russian example words (blue), respectively, that have been correctly (green) and incorrectly (red) aligned according to the gold standard.

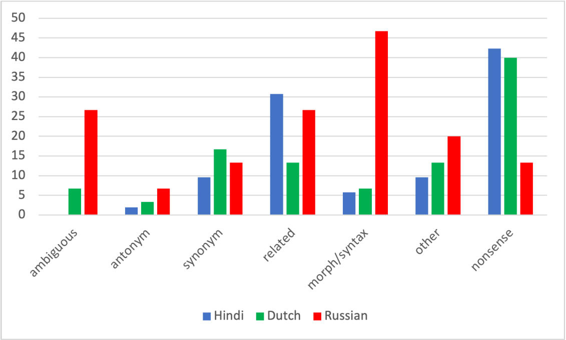


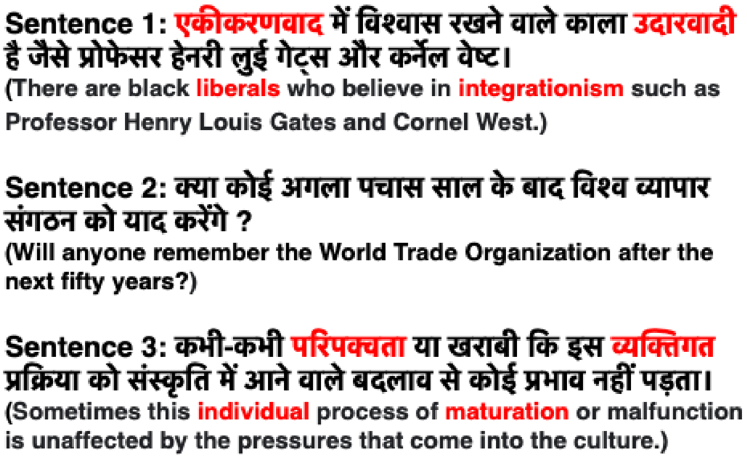
Figure 4: Distribution of error types per language (%)

as Chinese/Russian, and Thai/Hindi have near identicaly performance since they have more or less the same amounts of Wikipedia data. This really stresses the notion that the availability of monolingual resources is the primary bottleneck, while other reasons like language typology and sub-word tokenization might be secondary.

6. Conclusion

In this paper, we report cross-lingual transfer results for the extended anchor-based approach of Schuster et al (2019). Initially designed for ELMo embeddings, we evaluate the approach for the more recent BERT family of transformers for various monolingual and cross-lingual downstream tasks. We evaluate on one lower-resourced language, Hindi, while also presenting control results for three higher-resourced languages from a variety of language families, being Dutch, Russian and Chinese. It is clear from the experimental results that the language models and alignment methods perform worse for a lower-resourced language such as Hindi. Even though the method lags behind in lexical strength when compared to static word vectors, it beats a few baselines on the zero-shot XNLI task, but is unable to compete with the best approaches. We also attempted to analyze why the anchor approach, and most related cross-lingual approaches fail to perform for under-resourced languages. These results are in sync with works such as Wu et al. (2020), which demonstrate the under-representation of these languages even in a joint model like mBERT.

In future work, we would like to focus on developing high-quality evaluation sets for low-resourced languages so the state-of-the-art can be better assessed on tasks with a wider scope than NLI. Another interesting research direction is finding better transfer languages than English, since English is not an optimal pivot for most non-European languages (de Vries et



Sentence 1: एकीकरणवाद में विश्वास रखने वाले काला उदारवादी है जैसे प्रोफेसर हेनरी लुई गेट्स और कर्नेल वेष्ट।
(There are black liberals who believe in integrationism such as Professor Henry Louis Gates and Cornel West.)

Sentence 2: क्या कोई अगला पचास साल के बाद विश्व व्यापार संगठन को याद करेंगे ?
(Will anyone remember the World Trade Organization after the next fifty years?)

Sentence 3: कभी-कभी परिपक्वता या खराबी कि इस व्यक्तिगत प्रक्रिया को संस्कृति में आने वाले बदलाव से कोई प्रभाव नहीं पड़ता।
(Sometimes this individual process of maturation or malfunction is unaffected by the pressures that come into the culture.)

Figure 5: Examples from the XNLI Hindi test set for problematic sentences containing words (marked in red) with less than 10 occurrences in the Wikipedia validation set.

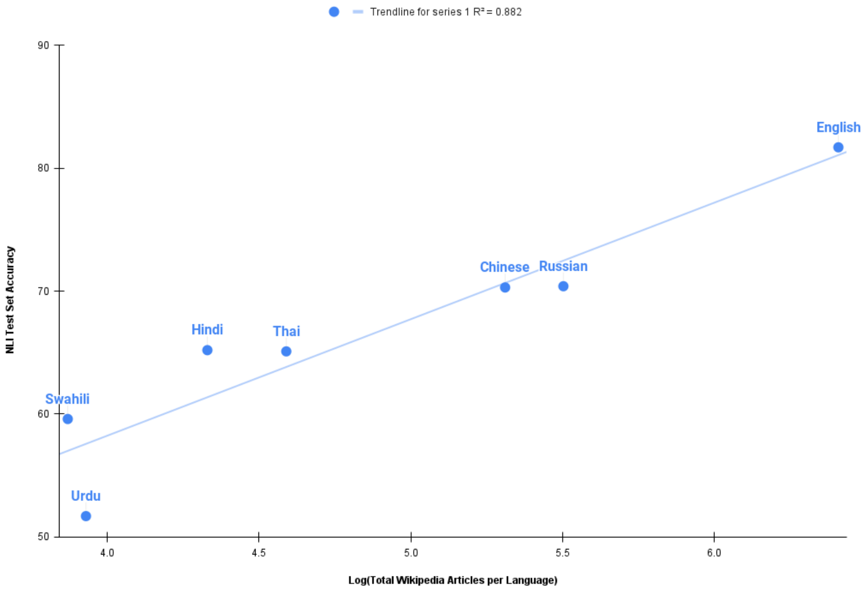


Figure 6: Plotting of different languages when taking the XNLI test accuracy (Y-axis) and number of Wikipedia pages (on a log scale) available for training (X-axis) into account.

al., 2022). Therefore, focusing on creating language-specific transformers jointly trained for a selection of closely related languages from the same language family could be a viable approach as well.
The extracted anchors for all 5 languages, modified MUSE dictionaries and other resources are made available at https://github.com/pranaydeeps/Vyaapak.

7. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. ACL.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July. Association for Computational Linguistics.

Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis.

Computer Speech Language, 28(1):56–75.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. *Advances in Neural Information Processing Systems*, pages 1853–1861.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland, June. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

de Vries, W., Nissim, M., and Wieling, M. (2022). Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online, May. Association for Computational Linguistics.

Delobelle, P., Winters, T., and Berendt, B. (2020). Robbert: a dutch roberta-based language model.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP 2016*, pages 1285–1295. Association for Computational Linguistics.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 976–983.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop Papers*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Rasooli, M., Farra, N., Radeva, A., Yu, F., and McKeown, K. (2018). Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1–2):143–165.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tran, K. M. (2020). From english to foreign languages: Transferring pre-trained language models. *CoRR*, abs/2002.07306.

Vulić, I., Ruder, S., and Søgaard, A. (2020). Are all good word vector spaces isomorphic?

Wang, Z., K, K., Mayhew, S., and Roth, D. (2020). Extending multilingual BERT to low-resource languages. *CoRR*, abs/2004.13640.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert? *CoRR*, abs/2005.09093.