Assessing the zero-shot technique on a cross-lingual transfer learning task

Tim Ottens
University of Amsterdam
t.ottens96@outlook.com

Thomas van Osch
University of Amsterdam
t.vanosch@hotmail.com

Abstract

Prior studies have shown the power of the zero-shot technique in a cross-lingual transfer learning setting. However, the effectiveness of the zero-shot approach compared to a pre-trained model is yet to be addressed. We propose such a comparison and evaluate on a masked language modeling task and a part-of-speech probing task. While a zero-shot approach achieves similar word-level accuracy on part-of-speech tagging compared to the native model, the transferred model lacks the contextual information required for masked language modeling.

1 Introduction

In recent years, pre-trained neural language models have significantly advanced the state of the art on a variety of natural language processing (NLP) tasks (Lample and Conneau, 2019; Radford et al., 2019; Devlin et al., 2018). Their powerful ability to fine-tune the model for downstream tasks, in which relatively small amounts of data may be available, has contributed to the furtherance of cross-lingual transfer learning. Particularly, zero-shot learning, which does not require labeled data of a target transfer language, sparked great interest in recent works (Artetxe and Schwenk, 2019; Eisenschlos et al., 2019). However, assessing the performance of the zero-shot method as opposed to a language model pre-trained on the target language is limited.

In this paper, we propose a comparison between a pre-trained BERT model on the Dutch language (Vries et al., 2019) and an English trained BERT model (Devlin et al., 2018) which employs Dutch embedding projections by a zero-shot approach. That is, the English model aims to understand the Dutch language without being exposed by Dutch labeled data.

Our findings suggest that based on the task of masked language modeling, the zero-shot fails to capture the richer context due to its word-level restrictions. On the word-level probing task, a zero-shot approach nears the performance of a pre-trained native model.

2 Related Work

Multilingual embeddings Ever since the fruitful development of word embeddings (Mikolov et al., 2013b), (Pennington et al., 2014) and the subsequent development in contextual word embeddings (Le and Mikolov, 2014), a wide variety of language models, such as CNN (Kim, 2014), LSTM (Peters et al., 2018) and Transformers (Radford et al., 2019; Raffel et al., 2019; Devlin et al., 2018) have set state-of-the-art results. However, these architectures increasingly grow in complexity and required training data, preventing the applicability of training from scratch on low-resource languages. Fortunately, novel language models are pre-trained and can be fine-tuned for downstream tasks, enabling cross-lingual learning for underrepresented languages.

Cross-lingual transfer LM Recently, pre-trained language models for cross-lingual transfer learning have been proposed, notably XLM, utilizing the BERT architecture (Lample and Conneau, 2019), LASER, employing a bi-directional LSTM model (Artetxe and Schwenk, 2019) and MultiFit (Eisenschlos et al., 2019) which is based on the ULMFiT model (Howard and Ruder, 2018). Generally, in this cross-lingual transfer learning domain two main techniques are employed. The zero-shot technique (Palatucci et al., 2009) is a favorable choice for truly low-resource language as no labeled data for the target language is required, however, the technique remains challenging on NLP tasks (Lample and Conneau, 2019; Artetxe et al., 2017; Artetxe and Schwenk, 2019; Eisenschlos et al., 2019). In contrast to the zero-shot

approach, in the few-shot setting the model does learn given labeled data from the target language.

3 Approach

In this section, the design choices for the zero-shot transfer learning comparison are discussed.

First, a model for comparison is selected. As BERT and its variants have shown convincing results on the zero-shot setting (Pires et al., 2019; Lample and Conneau, 2019), together with monolingual BERT models pre-trained specifically on a target language, we opted for the English pre-trained BERT (Devlin et al., 2018) and the Dutch pre-trained BERT model called BERTje (Vries et al., 2019). The zero-shot approach enables truly low-resource languages to be learned by a language model, as opposed to a supervised few-shot approach which requires labeled data of the target language. While such labeled data is being developed, the 6,000 languages globally are unlikely to be covered soon, suggesting the usefulness of the zero-shot setting we employ for the comparison.

3.1 Transformation matrix

Since in the zero-shot approach the target language cannot be processed directly by the pre-trained model on the source language, a transformation is required. This transformation is a mapping between the embeddings of the source language and the target language by either creating a shared vector space (Faruqui and Dyer, 2014) or transforming source embeddings to the target embedding space (Mikolov et al., 2013a). The latter approach commonly consists of a linear mapping that aims to minimize the distances between the source and target embedded words. The words are first translated by a bilingual dictionary and are then taken as input for the pre-trained model to retrieve the target embeddings. Examples of minimizing the mapping distance and thus finding the optimized transformation matrix are mean squared error (MSE) (Mikolov et al., 2013a), extending the MSE with an orthogonal constraint on the transformation matrix (Xing et al., 2015) or Canonical Correlation Analysis (Faruqui and Dyer, 2014). In this paper, the MSE with the orthogonal constraint is employed for finding the optimal transformation matrix due to its improved normalization, word length preservation and regularization in solving inconsistencies (Xing et al., 2015). On top of that, an exact closed-form solution can be estimated by Singular Value Decomposition (SVD). The mean squared error is formulated as the following optimization problem

$$\min_{W} \sum_{i} \|Wx_i - z_i\|^2 \tag{1}$$

where given a set of corresponding word embedding pairs $\{x_i, z_i\}_{i=1}^{n}$, with $x_i \in \mathbb{R}^{d_1}$ the embedding representation of word i in the source language and $z_i \in \mathbb{R}^{d_2}$ is the embedding representation of word i in the target language. W is the transformation matrix that aims to minimize the difference between $Wx_i - z_i$. The orthogonality constraint implies

$$\min_{\bar{W}} \|W - \bar{W}\| \text{ s.t. } \bar{W}^T\bar{W} = I \tag{2}$$

Moreover, as the target embeddings are sufficiently pre-trained, only the source embeddings are transformed without creating a joint embedding space.

3.2 Evaluation criteria

The performance of the two approaches will be assessed by multiple criteria to allow a fair comparison. Firstly, the masking task will be evaluated which aims to test the model's capability in predicting the masked words of a sentence. Criteria for these predictions will be the cross entropy loss and test perplexity. While this task may enlighten the contextual information in the embeddings, it does not give explicit insights into what aspects of language the models have learned. To this end, both approaches are probed on their ability to capture syntactic hierarchies in their internal representations (Hewitt and Manning, 2019). Specifically, a part-of-speech tagging task is performed together with a control task to account for the flexibility of this diagnostic classifier.

4 Experiments and Results

This section highlights the design choices made for the experiments and the results are discussed.

4.1 Experimental setup

Data The training data consists of a subset of the Dutch Wikipedia. The data is tagged on part-of-speech by the spaCy tagger (Honnibal and Montani, 2017). The bilingual dictionary is acquired from dict.cc (Hemetsberger, 2002).

Models Both models are cased, such that true case and accents are preserved which is favorable for the part-of-speech tagging task (Devlin et al.,

	Mask = 1		Mask = 20%		Mask = 50%	
Model	**CE**	**PP**	**CE**	**PP**	**CE**	**PP**
Dutch projection (Mikolov)	7.81 ± 0.02	61165 ± 5664	7.28 ± 0.01	10989 ± 457	7.04 ± 0.01	6356 ± 180
Dutch projection (Xing)	6.93 ± 0.02	24285 ± 357	6.85 ± 0.02	8325 ± 116	6.82 ± 0.01	7466 ± 3478
BERT (English)	7.36 ± 0.02	25505 ± 1183	7.32 ± 0.02	9997 ± 388	7.09 ± 0.02	5655 ± 177
BERTje (Dutch)	6.82 ± 0.01	1999 ± 117	6.73 ± 0.01	781 ± 17	8.33 ± 0.01	1511 ± 18

Table 1: Cross entropy (CE) and Perplexity (PP) on the masking task, with a masking of 1 token and 20% and 50% of the tokens in a sentence.

2018) and employ identical architectures. For the masking task, the BERT model is extended with a language modeling head. The models are implemented by the `transformer` library of Huggingface (Wolf et al., 2019).

Masking task In the masking task, n tokens are masked from the sentence and the complete embeddings of the sentence are taken as input to the language model. The output is a probability distribution over all tokens in the model. To evaluate the probability distribution x, a cross entropy loss $CE(x, t)$ is employed with target token t together with a perplexity metric that is defined as $PP(x) = 2^{CE(x,t)}$, to account for variation in the predictions loss. We report performance for different values for n.

Part-of-speech probing task The second experiment will test whether the projected embeddings incorporate word-level syntactic structures by predicting part-of-speech (POS) tags from the test dataset. To map the embeddings from their respective dimension to a probability distribution over POS-tags, a diagnostic classifier is employed which utilizes a 1-layer linear mapping. The predictions of the model are evaluated on accuracy. To further examine whether the accuracy is solely the result of the embeddings and not due to the learning capacity of the diagnostic classifier, a control task is performed (Hewitt and Liang, 2019). The control task is a prediction over randomized POS-tags for which every word in the vocabulary V has an identical mapping defined as $f_{control}(x_{1:T}) = f(C(x_1), C(x_2), ...C(x_T))$, where $C(x_i)$ maps the word token x_i to a POS-tag y_i (Hewitt and Liang, 2019).

4.2 Results

Table 1 shows the results of the masking task for the different models used in the experiments. Generally, the projection of Xing et al. (2015) achieves higher perplexity compared to the unconstrained optimization of Mikolov et al. (2013a), except when masking 50% of the tokens. In all cases, the native BERTje model outperforms the transferred BERT in terms of perplexity. Interestingly, increasing the number of masked tokens benefits the BERT model, but this trend is not analogous to BERTje. Masking more tokens may complicate the prediction task. Figures 1 and 2 shows the word distributions on the most probable predictions to examine the difference between the amount of masking and the behavior of the models.

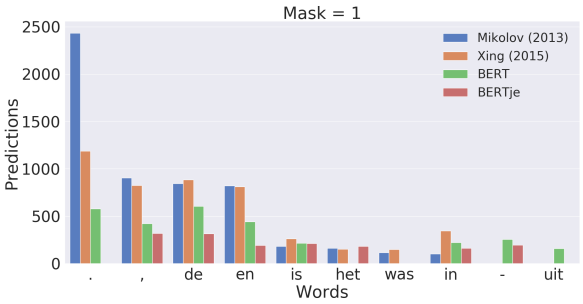

Figure 1: Top 8 predicted words on the masking task of one token.

In Figure 1 the word distributions are shown for masking one token in a sentence. The projection method mainly focuses on frequent punctuation, while the BERT models show a more uniform word distribution around words. Moreover, comparing BERT and BERTje shows the native BERTje to filter out punctuation more and the models differ in their most predicted words.

Results of increasing the number of masked tokens to 50% are shown in Figure 2 for the constrained projection and BERTje. The projection of Xing et al. (2015) focuses slightly more on real words and less on punctuation. The BERTje distribution shows a different trend compared to Figure 1. Less frequent words, together with the hyphen, account for the most predicted tokens, suggesting the drop in performance on this task. A possible explanation for this is the enhanced flexibility of the model in changing a masked sentence to its

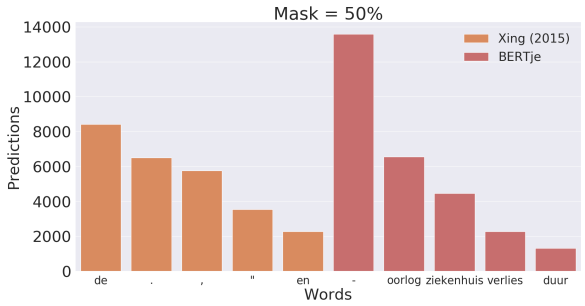

Figure 2: Top 5 predicted words on the masking task of 50% of the tokens.

Model	POS acc.	Control
Dutch projection (Mikolov)	0.95 ± 0.01	0.87 ± 0.00
Dutch projection (Xing)	0.95 ± 0.00	0.87 ± 0.00
BERT (English)	0.95 ± 0.01	0.89 ± 0.01
BERTje (Dutch)	0.95 ± 0.01	0.96 ± 0.01

Table 2: POS-tag prediction accuracy and control task accuracy on a total of 12 POS-tags

own liking.

Table 2 shows the result of the POS-tag prediction task. While each method achieves identical part-of-speech accuracy, suggesting no significant difference between the native and transferred model, the control task accuracy does differ across the models. Specifically, BERTje is outperformed by the transferred model, since the control task assigns random outputs, thus, the lower the better. A possible reason for this result is the relative sparse number of employed part-of-speech tags.

5 Discussion

5.1 Interpretation of the results

Overall, the transferred BERT model has shown to be capable of learning the Dutch language to an extent. In the masked modeling task, the difference between the native BERTje model and the English pre-trained BERT model with a zero-shot approach is not negligible. In this language modeling setting, having an understanding of the context of a masked token is essential and the relative naive linear mapping of the zero-shot approach in combination with the bilingual dictionary restricts the equalization of the transferred model. Moreover, it was shown the orthogonal constraint on the transformation matrix optimization to be fruitful for the zero-shot approach as it is more robust to punctuation due to its profitable gained properties proposed by Xing et al. (2015).

On the part-of-speech tagging task, the BERT model did achieve similar performance to the native BERTje, as shown in Table 2. The syntactic word-level understanding required for this task can be attained by a zero-shot approach. However, it should be noted that diversity of the annotated and predicted part-of-speech tags was limited, thus simplifying the prediction task. Additionally, the English and Dutch languages are syntactically very similar which further reduces the task difficulty. The native model did achieve the worst selectivity based on the high control task accuracy.

5.2 Limitations and future work

The results of the masked task on the different number of masked tokens do not unanimously favor the native model, suggesting careful interpretation. Further research may focus on finding an unambiguous explanation, for example, by elaborate probing on language modeling. Furthermore, a more diverse part-of-speech tagger may reveal more evident differences in the syntactic differences between the models, as the part-of-speech task only considers 12 tags. Besides part-of-speech probing, probing tasks on different linguistic properties, such as structural dependencies (Hewitt and Manning, 2019), may be of future work.

Lastly, improvements in zero-shot approaches prove to be of importance for languages that are underrepresented in labeled data. Developing projections that can utilize pre-trained models and their embeddings, which already have some form of syntactic properties, could give a boost to new related language models.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of

deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. 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 5702–5707, Hong Kong, China. Association for Computational Linguistics.

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

Paul Hemetsberger. 2002. English-dutch dictionary.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints*, page arXiv:1910.10683.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

A Appendix

A.1 Creating sentence embeddings

For the English BERT and Dutch BERTje model, the sentence tokens are taken as input to create individual word embeddings $x \in \mathbb{R}^{768}$, which correspond to the last hidden layer of both models. The projection models retrieve sentence embeddings from directly mapping words in the sentence to pre-trained Polyglot (Al-Rfou et al., 2013) embeddings $x \in \mathbb{R}^{64}$. Words that cannot be translated by the Dutch most frequent words from dict.cc (Hemetsberger, 2002) and are not in the Polyglot embeddings will be discarded.

A.2 Training the transformation matrix

The transformation matrix maps the Dutch embeddings to the English embedding space by applying the transformation matrix $W \in \mathbb{R}^{64\times768}$. This matrix is created with a Xavier uniform initialization and trained with Adam optimization and a learning rate of $1e^{-4}$. Since the loss is relatively small for training, a weighted scalar is applied with a value of 100. The transformation matrix is trained until early stopping with a maximum patience of 5 and an evaluation loss threshold of $1e^{-2}$. Differences between the optimization loss metrics are explained in Section 3.1

A.3 Masking task evaluation

For evaluating the projections, embeddings are mapped with the W transformation matrix to a dimension of 768. Subsequently, tokens are created by translation and taken as input to the corresponding tokenizer of BERT. The number of masked tokens in a sentence are selected according to the masking parameter k. The masked tokens are depicted with a special token for the model to recognize the tokens. Embeddings are retrieved from the BERT model and embeddings that do not correspond to the begin, end or masking token are replaced by the embeddings of the projection. These embeddings are then taken as input to the BertForMaskedLM model which outputs a prediction over all possible tokens for each word in the sentence. Predictions of the masked tokens are taken and a cross entropy loss is applied on the predicted probability distribution and the original token.

The English BERT model is evaluated likewise but without the projection. The native model BERTje does not require a projection of the data and creates embeddings in which the masked tokens are replaced by the special token of the tokenizer of BERTje. The BertForMaskedLM for Dutch embeddings will be pre-trained on the training data.

A.4 POS-tagging task evaluation

The part-of-speech tagging task is extended with a diagnostic classifier which consists of a 1-layer linear mapping from dimension 768 to 12 (number of POS-tags). These linear mappings will be trained with a cross entropy loss over the predicted POS-tags and an Adam optimizer with learning rate $1e^{-3}$. The classifier is trained in batches of 64-word embeddings and employs an early stopping mechanism with a maximum patience of 5 on the evaluation loss.