Exploring Multilingual Word Embedding Alignments in BERT Models: A Case Study of English and Norwegian

Pernille Aaby[1], Daniel Biermann[2], Anis Yazidi[1], Gustavo Borges Moreno e Mello[1], and Fabrizio Palumbo[1]

[1] Artificial Intelligence Lab (AI Lab), Institutt for informasjonsteknologi, Oslo Metropolitan University, Oslo, Norway
[2] Centre for Artificial Intelligence Research (CAIR) Department of ICT, University of Agder, Grimstad, Norway
fabrizio.palumbo@oslomet.no

Abstract. Contextual language models, such as transformers, can solve a wide range of language tasks ranging from text classification to question answering and machine translation. Like many deep learning models, the performance heavily depends on the quality and amount of data available for training. This poses a problem for low-resource languages, such as Norwegian, that can not provide the necessary amount of training data. In this article, we investigate the use of multilingual models as a step toward overcoming the data sparsity problem for minority languages. In detail, we study how words are represented by multilingual BERT models across two languages of our interest: English and Norwegian. Our analysis shows that multilingual models similarly encode English-Norwegian word pairs. The multilingual model automatically aligns semantics across languages without supervision. Additionally, our analysis also shows that embedding a word encodes information about the language to which it belongs. We, therefore, believe that in pre-trained multilingual modelsâĂŹ knowledge from one language can be transferred to another without direct supervision and help solve the data sparsity problem for minor languages.

Keywords: Natural Language Processing · Multilingual Bert · Word Alignment · Data Sparsity.

1 Introduction

Over recent years, the field of AI has made impressive progress regarding the performance of natural language processing tasks such as text classification, question answering, machine translation, or language generation. This progress is mainly driven by purely data-driven models such as transformers. To encode how words relate to their context, transformers are pre-trained on vast, unlabeled and mostly monolingual training corpora. This approach is powerful for languages such as English or Spanish, with an abundance of language resources consisting in raw text, labeled datasets, and benchmarks. However, when it comes to low-resource languages, such as Norwegian, the available language datasets are often limited. Unfortunately, the performance in such data-driven models and approaches heavily depends on the quality and amount of training data available. That is, good performance depends on high-quality datasets. At the

This is a post-peer-review, pre-copyedit version of the following conference proceeding: Aaby, P., Biermann, D., Yazidi, A., Mello, G.B.M.e., Palumbo, F. (2023). Exploring Multilingual Word Embedding Alignments in BERT Models: A Case Study of English and Norwegian. In: Bramer, M., Stahl, F. (eds) Artificial Intelligence XL. SGAI 2023. Lecture Notes in Computer Science(), vol 14381. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-47994-6_4

written time, there are 2181 matches for English datasets and only 67 for Norwegian datasets on huggingface.co[3]. More training data tend to improve the performance of language models [17,3]. Consequently, monolingual Norwegian language models will likely not achieve the same performance as monolingual English language models.

Most existing language models today have been trained on monolingual corpora [7,14], which do not benefit languages with sparse data availability. Isbister et al.[11] proposed an approach that translates the text from a low-resource language to a high-resource language. Then, it uses a state-of-the-art performing model trained on high-resource language to alleviate the data sparsity problem. However, recent work shows that specific multilingual language models manage to align words from different languages without learning from parallel data, which machine translation requires [4,15]. Therefore, we pose the questions:

- Can multilingual models relieve the need for monolingual models?
- Can knowledge from one language be transferred to another without parallel data?

In this article, we explore the similarities and dissimilarities between the word representations in English and Norwegian, using two multilingual language models. To this end, we use different methods from recent literature and combine them in a comprehensive study of the case of the English-Norwegian language pair.

To find similarities we evaluate word retrieval performance, from an English source vocabulary to a Norwegian target vocabulary. To find dissimilarities, we quantify the accuracy of retrieving the original language from the word representation. All methods are non-parametric and rely purely on vector proximity. The model architecture we have used is BERT (Bidirectional Encoder from Transformer) [7] since previous work has shown its capability to align words automatically [4,16].

We believe that this exploration can provide the research community with a better understanding of how the information of different languages manifests inside the word representations of multilingual models and ultimately help improve existing models and applications that suffer from data sparsity.

2 Related Work

2.1 Multilingual Word Retrieval

Mikolov et al. [22] noticed that the distribution of word embeddings in latent space showed similar characteristics across different languages. Motivated by the similarity of distributions, they hypothesized that they could align two distributions with word embeddings from two different languages to create a bilingual dictionary with word retrieval. Their technique relied on bilingual parallel corpora. Conneau et al. [6] showed that it was possible to align two-word embedding distributions from different languages without any supervision (parallel corpora). They utilized adversarial training to learn a linear mapping from the source to the target language, alleviating the need for parallel corpora.

[3] !https://huggingface.co/datasets Visited: 19.01.2023

2.2 Multilingual BERT

BERT is a transformer-based [30] model which improved state-of-the-art results on several NLP tasks at the time of release [7]. It improved on question-answering tasks like SQuAD v1.1 [25] and SQuAD v2.0 [24], and language understanding tasks like GLUE [31] and MutliNLI [33]. The model is trained on vast amounts of text corpora, the original English BERT used the English part of Wikipedia [7], but today it is being trained on bigger collections, even book collections from a whole library [13]. The model has been trained for several languages like French, Swedish, and Norwegian [20,19,14]. BERT can also be trained in several languages simultaneously to obtain multilingual understanding. mBERT is one of these models, and it is trained on Wikipedia corpus for 104 different languages, including English and Norwegian[4].

Notram, Norwegian Transformer Model, is a BERT model initialized from mBERT and further trained on mostly Norwegian book-corpus data [13]. Although the model is mainly trained on Norwegian corpus, after initialization, the authors estimate that a portion of 4% is English. The model scores high on Named Entity Recognition both for the Norwegian language and the English language.

Previous work [4,15] also shows that the semantics of two (and more) languages align automatically in BERT. So the model does not only represent two languages separately, but it is also able to encode connections between two languages through shared semantics of the words, without being trained on parallel data.

2.3 From Contextual to Static Embeddings

In order to benefit from previous benchmarks like SimLex999 [10], WordSim353 [1] and SimVerb3500 [9] that evaluate semantics, Bommasani et al. [2] distilled a set of static word embeddings from contextual word embeddings. This way the results could be compared to traditional word embeddings [21,23,12]. To create the static word embeddings from BERT they tried different aggregation and pooling strategies. The best-performing aggregation method was to take the average over several contexts, also referred to as AOC (Average Over Context). They also used *mean pooling*, taking the mean of all token representations over subtokens of a word in case a word consists of more than one token.

2.4 Probing BERT

Probing BERT has become a popular area of research to better justify its success and understand the model better so it is easier to improve the architecture [27]. It entails creating a simple classifier and using the features from the pre-trained model. If the simple classifier manages to solve the task, then we can assume that the necessary information is already within the features we extract.

From previous work, we know that BERT represents words with information about syntax and semantics [27]. Tenney et al. [28] discovered that BERT hierarchically learns information that corresponds to the traditional NLP (Natural Language Processing)

[4] !https://huggingface.co/bert-base-multilingual-cased

pipeline. Starting with local syntax structure such as POS tagging and parsing in the lower layers, while finding named entity recognition, semantic roles, and co-reference are information encoded later in the model in the respective order. Similar discoveries can be found in other works as well [16,29].

Naturally, since BERT is a contextual model representing a word based on not only itself but also the surrounding words, the question of whether one could distinguish different meanings of an ambiguous through the representation arose. In previous work [32,18] they find that ambiguous words divide different meanings into clusters from the contextual representation, although it is not always the same clusters as we would have defined from a human perspective.

3 Methods

Our analysis examines similarities and differences between word representations in two languages. Similarities are found through static word retrieval and differences through language detection. Our non-parametric method only relies on finding the most similar embedding(s) from a source word to a target collection. We used KNN (K- Nearest Neighbours) with cosine similarity to find the most similar vectors.

3.1 Static Word Retrieval

Following the work by Bommasani et al. [2] we created a static set of word embeddings by taking the AOC of several contextual embeddings for a term t. The contextual embedding for word t is obtained from a context $c_t \in C_t$, where each c_t is two sentences from the relevant language corpus.

$$s_t = \frac{1}{N_t} \sum_{n=1}^{N} w_{tn} \tag{1}$$

w_{tn} is the nth contextual embedding for the number of contexts $N_t = |C_t|$. For words that consist of more than one workpiece, we used mean pooling, taking the mean of all subtokens, to aggregate all token embeddings.

$$w_{tn} = \frac{1}{I_t} \sum_{i=1}^{I} p_{ti} \tag{2}$$

p_{ti} is the ith token in the word. We created static embeddings for all 13 intermediate representations from BERT, one after all the 12 stacked layers and the input layer. We aimed to retrieve a Norwegian target word from an English source word. The objective becomes, for each of the English word representations s_{i-en}, evaluate the cosine similarity to all the Norwegian word representations s_{j-no}, rank the similarities, and return the top(@) match(s). If a translation of the English word is one of the returned words, we achieved a correct word retrieval.

$$\textit{k-neighbours(i)} = \underset{j}{\operatorname{argmax}}\ sim(s_{i-en}, s_{j-no}) \tag{3}$$

$$y_i = \begin{cases} 1, & \textit{if k-neighbours(i)} \in \textit{translation}(s_{no}) \\ 0, & otherwise \end{cases} \tag{4}$$

$$\textit{accuracy static word retrieval} = \frac{1}{T}\sum_{i=1}^{T} y_i \tag{5}$$

T is the number of terms in the English vocabulary.

Liu et al. [15] test if word retrieval performance increases by doing a *mean shift*. Mean shift entails shifting from an English source word to be closer to a Norwegian target word by first subtracting the mean of all the English word embeddings and then adding the mean of all the Norwegian word embeddings. We define a language vector as the mean of all the static word embeddings in one vocabulary.

$$L_l = \frac{1}{T}\sum_{t=1}^{T} w_t \tag{6}$$

$l \in \{English, Norwegian\}$ and T is the number of words in each vocabulary. Mean shift:

$$s_{t,en->no} = s_{t,en} - L_{en} + L_{no} \tag{7}$$

L_{en} and L_{no} are language vectors for English and Norwegian respectively.

$$y_{i-l} = \begin{cases} 1, & \text{if} \quad sim(s_{i-l}, L_{en}) > sim(s_{i-l}, L_{no}) \quad and \quad l = en \\ 1, & \text{elif} \quad sim(s_{i-l}, L_{en}) < sim(s_{i-l}, L_{no}) \quad and \quad l = no \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\textit{accuracy language detection} = \frac{1}{2T}\sum_{i=1}^{T} y_{i-en} + \frac{1}{2T}\sum_{i=1}^{T} y_{i-no} \tag{9}$$

3.2 Language Detection

Motivated by the fact that words from the same language could be aggregated to a language vector, we asked the question:
Can we detect the language of a word based on the similarity to the language embeddings?
We detected the language by evaluating which language vector a word representation is most similar to.

3.3 Data

The Norwegian News Corpus[5] is used as the raw text corpora for the Norwegian part. We only used the part in Norwegian bokmÃěl (not nynorsk). The articles in the dataset

[5] !https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

are from multiple different newspapers, such as “VG”, “Aftenposten” and “Dagens nÃęringsliv” etc., collected from the years 1998 to 2019. We chose a set of contexts from the corpus for each word in our Norwegian vocabulary between 100 and 500. A context is defined as two sentences.

The vocabulary is restricted to only include the 50,000 most common words from the Norwegian News Corpus. In addition, we checked that the word is present in a Norwegian wordlist for BokmÃål [6].

To evaluate the word retrieval from English to Norwegian, we have used the English-Norwegian word benchmark from MUSE [7] [6]. We only used the word pairs, where the Norwegian word is in our top 50,000 vocabularies, and the English word is present in the Brown corpus [8] [8]. Some English words have more than one Norwegian word translation. We define a *correct word retrieval* as at least one match.

The Brown corpus gives the context sentences for the English word embedding vocabulary. The number of contexts for a word is the number of times a word stands in the Brown corpus but a maximum of 500 times. We only obtained static word embeddings for the words in the MUSE benchmark. The MUSE-filtered vocabulary ended up with approximately 12,000 English source words.

4 Results

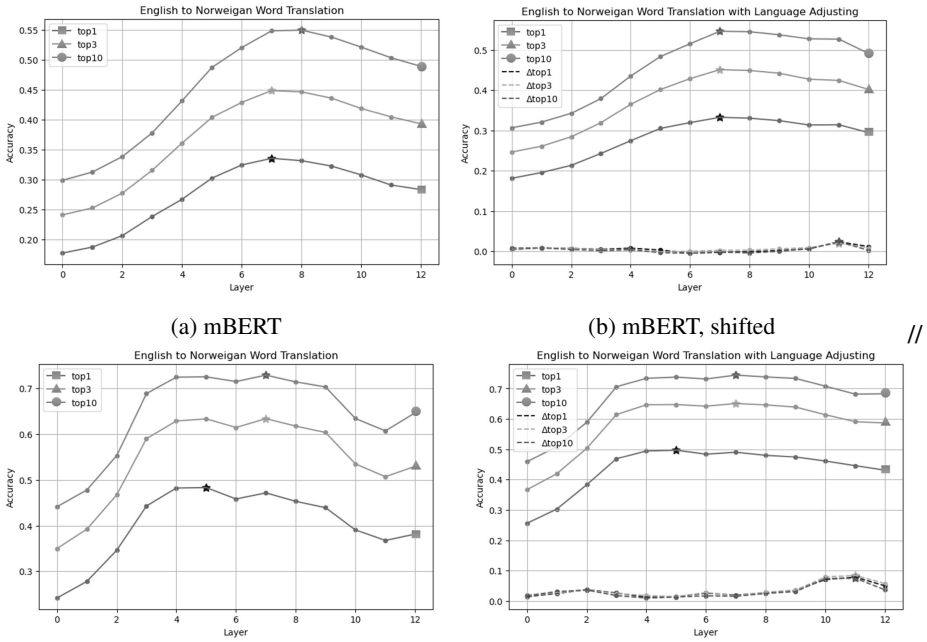
4.1 Static Word Retrieval

In Figure 1 we report the results of the English to Norwegian word retrieval using KNN and cosine similarity. We compare the performance of both mBERT (Figure 1a) and Notram (Figure 1c) for different numbers of top matches (@1,@3,@10). Notram achieved better accuracy than mBERT in general. The middle layers seem to perform best for both models, with Notram achieving around 50% at @1 match and more than 70% accuracy when using the @10 matches at layer 7. In addition, for the Notram model. we notice a dip in performance for layer 11. Overall, we argue that BERT models are capable of aligning semantics across English and Norwegian without using any supervised datasets with parallel sentences.

4.2 Static Word Retrieval with Mean Shift

Figure 1b and Figure 1d show the static word retrieval performances when adjusted with the mean shift. To illustrate the impact of the mean shift on the word retrieval performance better, the performance increase between the shifted and non-shifted model is depicted by the dashed lines. We can see that the overall influence of the mean shift on performance is relatively low across all layers. When mean shifting, the model retains word retrieval accuracy better from the middle layers to the subsequent layers than without the mean shift. The maximum word retrieval performance increase is reached in layer 11 for the Notram model, improving by 8% for K at @1, @3, and @10. Thus,

[6] !https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-23/
[7] !https://github.com/facebookresearch/MUSE
[8] !https://www.nltk.org/nltk_data/



(a) mBERT (b) mBERT, shifted

(c) Notram (d) Notram, shifted

Fig. 1: Static word retrieval performance from English to Norwegian with layer-wise performance accuracy with and without mean shift. The lower dashed lines depict the performance increase when using the mean shift. The star marker shows at which layer the performance peaked. Both models experience the highest performance increase in layer 11 for all chosen @matches.

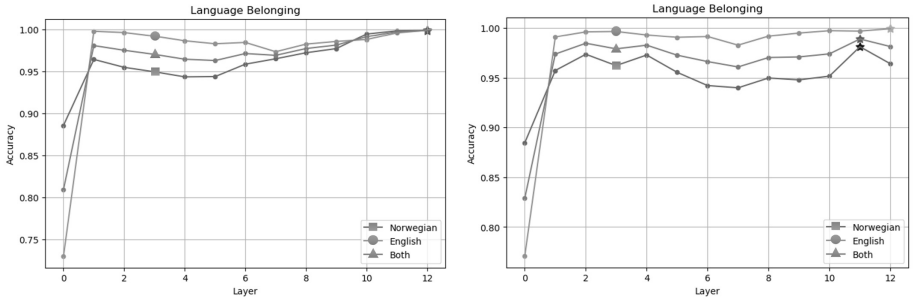
the mean shift seems to alleviate the cause of the performance dip seen before in later layers.

4.3 Language Detection

Figure 2 reports the results from the language detection experiment. The non-parametric method clearly shows that it is possible to find the language of a word using this method as the performance reaches almost 100% in the top-performing layer. The language detection accuracy reaches values above 95% for both models as soon as layer 1. This strongly indicates that the closest language vector can serve as a strong predictor for the language of the embedding.

4.4 Both Semantics and Language Properties can Cluster

For a more qualitative inspection of the word representations, Figure 3 illustrates both semantic alignment and language properties between English and Norwegian. The top



(a) mBERT: Language Belonging.

(b) Notram

Fig. 2: Layer wise language detection performance. The lighter line (circle) describes the prediction accuracy for the English vocabulary, the darkest line (square) describes the prediction accuracy for the Norwegian language and the line of intermediate shade (triangle) describes the combined prediction accuracy of language detection. The stars mark in which layer the performance peaks.

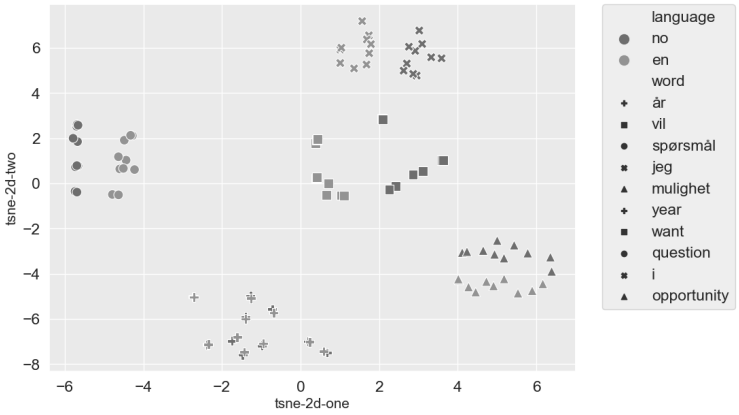
graph Figure 3a, inspired by previous work on semantic alignment in BERT [4], shows a plot comparing a set of 5 words in each English and Norwegian, respectively. The words were taken from the parallel corpus with sentences from riksrevisjonen[9][26]. We can observe that all word pairs are clustering together, indicating the semantic alignment of the word embeddings between the languages.

In the bottom graph Figure 3b we see two sets of 500 static word embeddings from each language. We can notice a clear clustering of the two languages. In both graphs, we reduce the embedding dimension to two dimensions with the t-SNE method. This further solidifies that BERT models are able to align semantics across English and Norwegian without using any supervised data

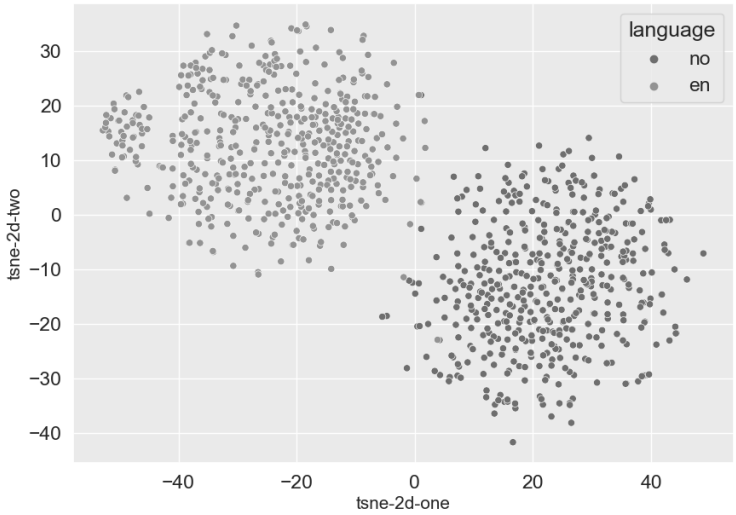
5 Discussion

Our analysis shows that layers 5-9 (middle layers) have the highest accuracy on static word retrieval. This result is in line with previous work on semantic similarity [5]. We argue that the best-performing layers in semantic similarity will also be the best-performing layers in semantic alignment between two languages. Although we observe a clear separation between languages in the word representation space, the mean shift method did not significantly impact the word retrieval accuracy. In layer 11, the accuracy does increase by around 8% in the Notram model. However, in the best performing layer of the same model, layer 7 (or 5), the increase is only around 1%. We consider this a slight change since the accuracy at @1 is around 50%. Overall, the word retrieval

[9] !https://www.elrc-share.eu/repository/browse/bilingual-english-norwegian-parallel-corpus-from-the-office-of-the-auditor-general-riksrevisjonen-website/a5d2470201e311e9b7d400155d0267060fffdc9258a741659ce9e52ef15a7c26/



(a) Visualizing contextual word embeddings for word pairs in English and Norwegian. Contextual embeddings taken from layer 8 of the Notram model and the English contextual embeddings have experienced a mean shift. The word embeddings are reduced to 2D with t-SNE. The darker colored markers show contextual word embeddings for Norwegian while the lighter color show contextual embedding for English. Each word pair has its own marker.



(b) 500 random words from Norwegian and English vocabulary respectively. The static word embeddings are from Notram layer 12. The word embeddings are reduced to 2D with t-SNE. Lighter points correspond to English embeddings and darkest points correspond to Norwegian embeddings.

Fig. 3: Visualizing static and contextual word embeddings from BERT

results suggest that the hypothesis that translating one language to another in the word representation space by looking at the closest matched embedding of the other language is a promising approach. Though, the low impact of the mean shift indicates that the translation from one language to another is not as simple as shifting the embedding by a simple mean language vector. This warrants further investigation into better methods to create language vector representations that might improve the impact of such a language vector shift. Nevertheless, the language vectors from the mean shift analysis remain strong predictors for identifying the language of an embedding as can be seen by the strong performance results of our language detection analysis.

It is noteworthy that static word retrieval does not deal with ambiguous words. Both language vocabularies most likely contain words with multiple meanings, leading to a conflation of meaning in the embedding. Conflated meanings most likely affected word retrieval since the English and Norwegian corpus do not provide the same contexts, and a word representation can be conflated with different meanings depending on the text corpus. In addition, words within each language can have different meanings. Therefore, an ambiguous word can often be detected because it will translate to different words in another language depending on the context. To deal with this downside, one would have to include a more nuanced analysis of either sense or a word pair from the same context. We believe that ambiguous words have a negative impact on accuracy as we could observe significantly better results when considering @3 and @10 nearest neighbours, with an increase of more than a 20% going from @1 to @10.

Norwegian is a language that borrows many words and phrases from English. It can be single words like "skateboard" or whole phrases like movie titles. Even though we filtered out sentences detected as English from the Norwegian text corpus, single words and smaller phrases may have been hard to remove. The effect could be English *noise* in the Norwegian part of the corpus and hence an effect in language detection. mBERT outperforms Notram in the subsequent layers of the model in detecting the correct language, and it achieves close to 100% accuracy. However, we question if the accuracy is this high because there might exist English noise in the Norwegian corpus, which would mean that the accuracy should not be 100%. A better evaluation dataset could be used to inspect this effect further.

6 Conclusion

In this exploratory analysis, we have shown that BERT's word representations automatically align semantics across English and Norwegian. We showed this with an accuracy of 50% for @1 nearest neighbor and an accuracy of more than 70 % for @10 nearest neighbor on the word retrieval task. In addition, we found that language is encoded in the word representation: We could detect the correct language of a word, with close to 100% accuracy, only by looking at its proximity to the two language vectors for English and Norwegian, respectively. We demonstrate that the model can align semantics and learn language properties by training on only raw text data (no parallel sentences).

We believe that the combination of automatic language detection and word retrieval between language embeddings allows for knowledge to be transferred between languages, ultimately helping alleviate the data sparsity problem in low-resource lan-

guages, such as Norwegian. While our results show promising tendencies, further investigations into reaching higher word retrieval accuracies and better aligning language vectors are warranted to make this approach reliable. We hope that our findings motivate new ways of using multilingual models and inspire more research in training and investigating multilingual models for low-resource languages.

References

1. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A.: A study on similarity and relatedness using distributional and wordnet-based approaches (2009)
2. Bommasani, R., Davis, K., Cardie, C.: Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4758–4781 (2020)
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165 (2020)
4. Cao, S., Kitaev, N., Klein, D.: Multilingual alignment of contextual word representations. arXiv preprint arXiv:2002.03518 (2020)
5. Chronis, G., Erk, K.: When is a bishop not like a rook? When itâĂŹs like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In: Proceedings of the 24th Conference on Computational Natural Language Learning. pp. 227–244 (2020)
6. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**(Mlm), 4171–4186 (2019)
8. Francis, W.N., Kucera, H.: Brown corpus manual. Letters to the Editor **5**(2), 7 (1979)
9. Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A.: Simverb-3500: A large-scale evaluation set of verb similarity. arXiv preprint arXiv:1608.00869 (2016)
10. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics **41**(4), 665–695 (2015)
11. Isbister, T., Carlsson, F., Sahlgren, M.: Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead? arXiv preprint arXiv:2104.10441 (2021)
12. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. vol. 2 (2017). https://doi.org/10.18653/v1/e17-2068
13. Kummervold, P.E., la Rosa, J., Wetjen, F., Brygfjeld, S.A.: Operationalizing a national digital library: The case for a norwegian transformer model. arXiv preprint arXiv:2104.09617 (2021)
14. Kutuzov, A., Barnes, J., Velldal, E., Øvrelid, L., Oepen, S.: Large-scale contextualised language modelling for norwegian. arXiv preprint arXiv:2104.06546 (2021)
15. Liu, C.L., Hsu, T.Y., Chuang, Y.S., Lee, H.Y.: A study of cross-lingual ability and language-specific information in multilingual BERT. arXiv preprint arXiv:2004.09205 (2020)

16. Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., Smith, N.A.: Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:1903.08855 (2019)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR **abs/1907.1** (2019), http://arxiv.org/abs/1907.11692
18. Loureiro, D., Rezaee, K., Pilehvar, M.T., Camacho-Collados, J.: Analysis and Evaluation of Language Models for Word Sense Disambiguation. Computational Linguistics pp. 1–55 (2021)
19. Malmsten, M., Börjeson, L., Haffenden, C.: Playing with Words at the National Library of Sweden–Making a Swedish BERT. arXiv preprint arXiv:2007.01658 (2020)
20. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de La Clergerie, Ã.V., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. arXiv preprint arXiv:1911.03894 (2019)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings **ICLR 2013**, 1–12 (2013), https://arxiv.org/abs/1301.3781
22. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
23. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2014). https://doi.org/10.3115/v1/d14-1162
24. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822 (2018)
25. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuad: 100,000+ questions for machine comprehension of text. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings (ii), 2383–2392 (2016). https://doi.org/10.18653/v1/d16-1264
26. Riksrevisjonen: Bilingual English-Norwegian parallel corpus from the Office of the Auditor General (Riksrevisjonen) website âĂŞ ELRC-SHARE (2018), https://www.elrc-share.eu/repository/browse/bilingual-english-norwegian-parallel-corpus-from-the-office-of-the-auditor-general-riksrevisjonen-website/a5d2470201e311e9b7d400155d0267060fffdc9258a741659ce9e52ef15a7c26/
27. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2020)
28. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950 (2019)
29. Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Van Durme, B., Bowman, S.R., Das, D., others: What do you learn from context? probing for sentence structure in contextualized word representations. arXiv preprint arXiv:1905.06316 (2019)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Å., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **2017-Decem**(Nips), 5999–6009 (2017)
31. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
32. Wiedemann, G., Remus, S., Chawla, A., Biemann, C.: Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. arXiv preprint arXiv:1909.10430 (2019)
33. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)