Word Alignment with Multilingual BERT : From Subwords to Lexicons

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

October 2025

1 Introduction

In this lab, we will explore how to extract a bilingual lexicon (a list of word translation pairs linking a source language to a target language) directly from the internal representations of a "large" multilingual language model. Specifically, we will rely on `mBERT`, a variant of BERT pretrained on Wikipedia in over 100 languages (Devlin et al. 2019).

As we discussed at (too) great length in class, one of the surprising findings about `mBERT` is its ability to perform cross-lingual transfer without any explicit alignment objective or parallel data (Pires, Schlinger et Garrette 2019 ; Wu et Dredze 2019). Although trained solely with a masked language modelling objective on monolingual corpora, mBERT appears to learn a shared semantic space across languages. Words that are translations of each other tend to occupy neighbouring regions in the representation space, even when the languages do not share scripts or subwords.

Because of this emergent alignment, it is possible to recover word-level translation pairs by comparing the contextualised embeddings produced by `mBERT` for parallel sentences. In this lab, we will experiment with three different alignment strategies :

- Direct argmax alignment, where each source word is linked to the target word with the most similar embedding.
- Competitive linking (via the Hungarian algorithm), which enforces one-to-one alignments.
- Canonical Correlation Analysis (CCA), which learns linear projections to better align the representation spaces of two languages (Cao, Kitaev et Klein 2020).

2 Corpora

To carry out this lab, we will rely on a range of multilingual resources. First, we require parallel corpora (collections of sentence pairs that are translations of each other) so that we can compare word representations across languages. Well-known examples include the No Language Left Behind (NLLB) dataset (Team et al. 2022), which provides large-scale, high-quality translations across more than 200 languages. In addition, we will use bilingual lexicons, that is, precompiled lists of word translation pairs such as those available in MUSE (Lample et al. 2018)[1] or PanLex[2] (Kamholz, Pool et Colowick 2014) These lexicons serve both as supervision signals (for instance when applying Canonical Correlation Analysis) and as gold standards to evaluate the accuracy of the alignments we obtain.

3 Word-level alignment methods

We (or to be more precise "you" ☺️) will experiment with three different strategies for extracting word-level alignments from multilingual representations produced by **mBERT**. Each method illustrates a distinct approach to exploiting cross-lingual similarity, and each raises important considerations with respect to tokenisation, similarity metrics, and projection techniques.

Direct argmax alignment The most straightforward method is to align each source-language word with the target-language word whose embedding is most similar. The procedure is as follows :

1. encode a parallel sentence pair with `mBERT` ;
2. extract embeddings for each word, taking care to reconstruct word-level representations from subwords (e.g., by averaging or taking the first subtoken) ;
3. compute pairwise similarities between source and target words, typically using cosine similarity, as it is scale-invariant and widely adopted ;
4. for each source word, select the target word with the highest similarity.

This method is simple and efficient but often produces many-to-one alignments and is sensitive to errors arising from subword tokenisation.

Competitive linking (one-to-one alignment) To overcome the problem of many-to-one correspondences, we can enforce one-to-one alignments between words. This can be formulated as a bipartite matching problem : construct a similarity matrix between all source and target words, then apply an assignment algorithm such as the Hungarian

[1] https://github.com/facebookresearch/MUSE
[2] https://panlex.org

(Kuhn–Munkres) algorithm to select the matching that maximises the overall similarity while guaranteeing that each word is linked to at most one counterpart. This “competitive linking” produces more coherent alignments, although its quality still depends on tokenisation choices and the similarity measure.

Canonical Correlation Analysis (CCA) The third method improves alignment by learning projections of the representation spaces before similarity is computed. Canonical Correlation Analysis (CCA) learns two linear mappings that maximise the correlation between embeddings of known translation pairs. In practice, one gathers a small bilingual lexicon, extracts word embeddings from `mBERT` (again carefully aggregating subwords), and organises them into two matrices, one per language. CCA then computes projection matrices that map both sets of embeddings into a shared space where translation equivalents are highly correlated. Once trained, new embeddings can be projected into this space before applying cosine similarity. Compared to the previous methods, CCA explicitly calibrates the embedding spaces, often yielding higher-quality alignments, provided that the seed lexicon is representative.

4 Implementation

1. Select around ten languages, ensuring that you include both languages covered during `mBERT` pre-training (e.g. French, German, Arabic, Chinese) and languages that are less represented or absent from its training data (e.g. Yoruba, Hausa, Quechua).
2. For each selected language, extract a parallel corpus with English (for instance approximately 5,000 sentence pairs). Suitable resources include *Tatoeba*, *JW300*, or the Bible corpus. Then, obtain the `mBERT` representations for all these sentences, aggregating subword embeddings in order to reconstruct word-level vectors.
3. Implement the first alignment method (argmax over cosine similarities) and evaluate its performance. Discuss the most appropriate evaluation metrics and how they should be applied (on all words in the sentence, or only on a pre-identified subset of words, etc.). Observe whether performance is consistent across languages or whether some languages benefit more than others.
4. Investigate whether enforcing a one-to-one word-level constraint (using the Hungarian algorithm for *competitive linking*) improves alignment performance compared with direct argmax alignment.
5. Next, implement the method based on Canonical Correlation Analysis (CCA). As usual, consulting the `scikit-learn` documentation is advisable. Two strategies should be tested : (i) learning a separate CCA for each language–English pair ; (ii) reusing projection matrices learned for one language and applying them to a different language.

6. Finally, consider whether the learned projection matrices and/or the observed alignment performance can be used to measure a form of *distance* between languages. In particular, can we quantify how close each language appears to English in the representation space induced by `mBERT` ?

Références

Cao, Steven, Nikita Kitaev et Dan Klein (2020). "Multilingual Alignment of Contextual Word Representations". In : *International Conference on Learning Representations*. url : https://openreview.net/forum?id=r1xCMyBtPS.

Devlin, Jacob et al. (juin 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)*. Sous la dir. de Jill Burstein, Christy Doran et Thamar Solorio. Minneapolis, Minnesota : Association for Computational Linguistics, p. 4171-4186. doi : 10.18653/v1/N19-1423. url : https://aclanthology.org/N19-1423/.

Kamholz, David, Jonathan Pool et Susan Colowick (mai 2014). "PanLex : Building a Resource for Panlingual Lexical Translation". In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Sous la dir. de Nicoletta Calzolari et al. Reykjavik, Iceland : European Language Resources Association (ELRA), p. 3145-3150. url : https://aclanthology.org/L14-1023/.

Lample, Guillaume et al. (2018). "Word translation without parallel data". In : *International Conference on Learning Representations*. url : https://openreview.net/forum?id=H196sainb.

Pires, Telmo, Eva Schlinger et Dan Garrette (juill. 2019). "How Multilingual is Multilingual BERT ?" In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Sous la dir. d'Anna Korhonen, David Traum et Lluís Màrquez. Florence, Italy : Association for Computational Linguistics, p. 4996-5001. doi : 10.18653/v1/P19-1493. url : https://aclanthology.org/P19-1493/.

Team, NLLB et al. (2022). *No Language Left Behind : Scaling Human-Centered Machine Translation*. arXiv : 2207.04672 [cs.CL]. url : https://arxiv.org/abs/2207.04672.

Wu, Shijie et Mark Dredze (nov. 2019). "Beto, Bentz, Becas : The Surprising Cross-Lingual Effectiveness of BERT". 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)*. Sous la dir. de Kentaro Inui et al. Hong Kong, China : Association for Computational Linguistics, p. 833-844. doi : 10.18653/v1/D19-1077. url : https://aclanthology.org/D19-1077/.