Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

Anchors Selection for Cross-lingual Embedding Alignment through Time

Filippo Pallucchini
Dept of Statistics and Quantitative Methods, University of Milan-Bicocca
filippo.pallucchini@unimib.it

Abstract

In recent years, vector representations of words have proven to be extremely useful across a wide range of NLP applications. Because of the broad interest in the topic, it became essential answering the following question: is it possible to align different embeddings, to compare terms belonging to different vector spaces and their relations? While embedding alignment (*EA*) received considerable attention in literature, how to find the best anchors is still an open problem; in this paper, we propose an unsupervised, automatic method to select words belonging to different corpora that are close from a semantic point of view, and can be used as anchors for aligning their respective embedding spaces.

1 Introduction and Motivation

Learning widely applicable representations of words has been an active area of research for decades; for this purpose, have been used a wide range of methods including co-occurrence matrix factorization [Pennington *et al.*, 2014], neural networks [Mikolov *et al.*, 2013] and transformers [Devlin *et al.*, 2018]. Because of the broad interest in the topic, it became essential answering to this question: is it possible to make two different word vector models consistent with each other, in order to transfer lexical and semantic knowledge across different corpora? Indeed, most of the embedding models, even if trained on the same corpus, produce misaligned vector spaces. In this research, we focus on the alignment of cross-lingual (*CL*) embeddings over time. The possibility to align *CL* word embeddings (*WEs*) spaces is appealing for two reasons: first, this enable to compare the meaning of words across languages, which is key to bilingual lexicon induction, machine translation, *CL* information retrieval; second, *CL WEs* enable model transfer between languages, e.g., between resource-rich and low-resource languages, by providing a common representation space [Ruder *et al.*, 2019]. Most methods to learn bilingual *WEs* rely on large parallel corpora, which are difficult to obtain for most language pairs [Artetxe *et al.*, 2017]. We focus our attention over those methods of alignment that rely on *anchors* (*seed lexicon*), i.e terms that have the same meaning in both corpora. They are used as reference points for mapping the embeddings from one space into the other by minimizing their distances.

1.1 Objectives and Research Questions

In literature, existing techniques to perform anchor selections suffer from some limitations; for instance, some methods need labeled corpus, other methods rely on non-general assumptions that limit or weaken the applications of the alignment (e.g. assume that some specific terms have the same meaning in different corpora). Our scope is to build a model, domain and time independent, to select anchors for the alignment of two different embedding spaces (i) automatically (ii) without the need of labeled corpus (iii) considering lexical and semantic knowledge of the embeddings.

An Inspiring Example in the Labour Market. Consider two embeddings generated from Job Ads (*JAs*), UK and IT. Word vectors of occupations/skills embed semantic and lexical information of the two labor markets and in the time the *JAs* have been posted. However, it might happen that some terms change their use over time or have a different meaning in the two countries. E.g., the occupation called *Data Scientist* in UK, might require skills and mansions that are requested to a *Data Engineer* in IT. Or it might that the use made by recruiters of the term *CISO* now is not the same of 5 years ago. How can we compare this two labour markets, or *JAs* in different years, using their embeddings? It would be partial to select common general words as anchors for the alignment, for the reasons illustrated in Sec 1. So, we propose to select terms that are semantically and lexically similar in the two spaces and use them as Seed Lexicon for the alignment.

1.2 State of the Art

As mentioned before, it is becoming increasingly important to track and detect linguistic shifts in the usage of words over time and across different languages and domains. *CL*/Cross-domain correspondences play key roles in tasks ranging from machine translation to transfer learning. These tasks can be achieved comparing *WEs* trained in different temporal periods/languages. Mikolov et al. [Mikolov *et al.*, 2013] and Kulkarni et al. [Kulkarni *et al.*, 2015] built methods based on the idea that some concepts in two different languages have similar geometric arrangements (e.g. animal and numbers).

The reason is that as all common languages share concepts that are grounded in the real world (such as that cat is an animal smaller than a dog), there is often a strong similarity between the vector spaces. So, they do not select specific anchors for the alignment but use just the most frequent words in both corpora. This hp carries with it some limitations; indeed, it may not perform well if applied with two corpora that have intrinsically a strong difference (e.g. corpora of different domains or two different languages with different cultures). Another common approach is to rely on shared words and cognates [Peirsman and Padó, 2010], eliminating the need for bilingual data in practice. E.g., Artetxe's method [Artetxe *et al.*, 2017] exploits the structural similarity of embedding spaces, creating a numeral dictionary, consisting of words matching the [0-9]+ regular expression on both vocabularies (e.g. 1-1, 1992-1992...). This idea assumes that these terms have the same meaning in both corpora and can guide the alignment process. However, generic words like numerals could have completely different usage in different corpora (e.g. 1996 could be a year but also the product's price).

2 Proposed Method

To the top of our knowledge, so far papers that propose unsupervised methods for anchors selection carry some limitations; in particular, they do not consider the semantic similarity of words between corpora. We decided to develop a method that automatically finds the best anchors for embeddings alignment selecting words that do not present a shift in the meaning. To this aim, we compute a score of semantic similarity for each pair of words in common among the two corpora (for *CL* alignment the aid of a Translator is needed; for our experiments we are using on-line Google Translate). Given two corpora (e.g. in different languages) and their respective vocabularies I and J, let X and Y denote their two respective embedding matrices so that X_i corresponds to the ith source language word vector and Y_j corresponds to the jth target language word vector. For each X_i and Y_j we extract the set of k-most similar words in the respective embedding, M_{Xi} and N_{Yj} so that $m_i = \{m \in M_{Xi}\}$ and $n_j = \{n \in N_{Yj}\}$; then, we compute a similarity score that consider m_i and n_j, $score_{ij} = f(m_i, n_j)$. This score allows us to have a criteria for the choice of best seed lexicon. In particular, we are going to select d pairs of i and j with highest $score_{ij}$ considering a dispersion constraint ϵ in order to avoid the selection of anchors concentrated in limited parts of the vector space, so that we create a dictionary $D = \{\delta \in \mathcal{I}, \delta \in \mathcal{I} \mid best(score_\delta)\ and\ cosinesim(m_\delta, n_\gamma)\ < \epsilon, \gamma \neq \delta\}$. After this process it is possible to find a transformation matrix W such that WX_δ approximates Y_δ through a simple optimization problem: $\min_{\mathbf{W}} \sum_{\delta=1}^{d} ||WX_\delta - Y_\delta||^2$. Therefore, we define two words as similar if they have not just the same morphology (translated or not) but also considering all words closest in the embedding space; these would assure the selection of anchors that have the highest semantic similarity in both corpora. Each word or character could have a different meaning in a specific domain or language, so, it is not possible to define first a list of words to use as anchors but it is necessary to build a method effective for all possibilities. To evaluate our method we performed bilingual lexicon extraction, which measures the accuracy of the induced dictionary in comparison to a well-known gold standard, in our case the dataset proposed by [Dinu *et al.*, 2014]. Preliminary experiments showed promising performances, outperforming the state-of-the-art method proposed in [Artetxe *et al.*, 2017].

3 Expected Contributions to the Community

We propose a new approach to automatically select the best anchors for *EA*. The idea is to create a method applicable to every embedding model, time and domain independent. We believe that this work would have implications for the fields of Semantic Search and the recently burgeoning field of Internet Linguistics. Another interesting application could be the building of a translator; once the alignment is performed we can translate any word of the monolingual corpora by projecting its vector representation from the source language space to the target language one. Future works are aimed at optimizing this approach and extending it to the world of contextual embeddings; indeed, they could yield richer representations of meaning compared to their static counterparts but aligning them poses a big challenge due to their dynamic nature.

Acknowledgments

This work has been achieved also thanks to prof. Fabio Mercorio, prof. Mario Mezzanzanica and Dr. Lorenzo Malandri.

References

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, pages 451–462, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[Dinu *et al.*, 2014] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv:1412.6568*, 2014.

[Kulkarni *et al.*, 2015] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *WWW '15*, pages 625–635, 2015.

[Mikolov *et al.*, 2013] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv*, 2013.

[Peirsman and Padó, 2010] Yves Peirsman and Sebastian Padó. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *NAACL*, pages 921–929, 2010.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Ruder *et al.*, 2019] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *JAIR*, 65:569–631, 2019.