Evaluating historical word embeddings: strategies, challenges and pitfalls

Oksana Dereza, Theodorus Fransen and John P. McCrae
University of Galway, Insight Centre for Data Analytics

When it comes to the quantitative evaluation of word embeddings, there are two main strategies: extrinsic, i.e. using pre-trained embeddings as input vectors in a downstream ML task, such as language modelling, and intrinsic, i.e. through analogy and similarity tasks that require special datasets (Bakarov, 2018).

Extrinsic evaluation

Language modelling seems to be the easiest way to evaluate historical word embeddings, since it is language independent, scalable and does not require dataset creation. Hypothetically, using pre-trained embeddings must lower the perplexity of a language model, even if these embeddings were trained on a different period of the same language. However, language modelling, as well as the majority of modern NLP tasks, is not very relevant to historical linguistics, so we might want to find a better downstream task or turn to intrinsic evaluation.

Intrinsic evaluation

There are two major tasks used for intrinsic evaluation of word embeddings: similarity and analogy. The **similarity task** consists in comparing similarity scores of two words yielded by an embedding model to those calculated based on experts' judgment. We did not explore this option, because it requires too much manual work by definition. The **analogy task** is simply asking an embedding model "What is to **a′** as **b** is to **b′** ?", and expecting **a** as an answer. Analogy datasets can be created automatically or semi-automatically if there exists a comprehensive historical dictionary of a language in question in machine readable format or a WordNet.

Traditionally, analogy datasets are based on pairwise semantic proportion and therefore every question has a single correct answer. Given the high level of variation in historical languages, such a strict definition of a correct answer seems unjustified. Therefore, in our Early Irish analogy dataset we follow the authors of BATS (Gladkova et al., 2016) providing several correct answers for each analogy question and evaluating the performance with set-based metrics, such as an average of vector offset over multiple pairs (3CosAvg).

Our dataset consists of 4 parts: morphological variation and spelling variation subsets were automatically extracted from eDIL (eDIL, 2019), while synonym and antonym subsets are translations of correspondent BATS parts proofread by 4 expert evaluators. However, the scores that Early Irish embedding models achieved on the analogy dataset were low enough to be statistically insignificant. Such a failure may be a result of the following problems:

The highest inter-annotator agreement score (Cohen's kappa) between experts was 0.339, which reflects the level of disagreement in the field of historical Irish linguistics. It concerns such fundamental questions as "What is a word? Where does it begin and end? What is a normalised spelling of a word at a particular stage of the language history?", which was discussed in (Doyle et al., 2018) and (Doyle et al., 2019) regarding tokenisation. It is arguable that it might be true for historical linguistics in general.

There is a lack of standardisation in different resources for the same historical language. For example, ~65% of morphological and spelling variation subsets, retrieved from eDIL, were not present in the whole Early Irish corpus retrieved from CELT (CELT, 1997), on which the biggest model was trained. As for synonym and antonym subsets, ~30% are missing in the corpus. Although our embedding models used subword information and were able to handle unknown words, such a discrepancy between the corpus, on which they were trained, and the historical dictionary, which became the source for the evaluation dataset, seriously affected the performance. This discrepancy originates from different linguistic views and editorial policies used by different text editors, publishers and resource developers throughout time.

References

Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. ArXiv:1801.09536 [Cs]. http://arxiv.org/abs/1801.09536

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. ArXiv:1607.04606 [Cs]. http://arxiv.org/abs/1607.04606

CELT: Corpus of Electronic Texts. (1997). University College Cork. http://www.ucc.ie/celt

Doyle, A., McCrae, J. P., & Downey, C. (2018). Preservation of Original Orthography in the Construction of an Old Irish Corpus. Proceedings of the LREC 2018 Workshop "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age", 67–70.

Doyle, A., McCrae, J. P., & Downey, C. (2019). A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. Proceedings of the Celtic Language Technology Workshop, 70–79. https://www.aclweb.org/anthology/W19-6910

EDIL 2019: An Electronic Dictionary of the Irish Language, based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913-1976). (2019). www.dil.ie

Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. Proceedings of the NAACL Student Research Workshop, 8–15. https://doi.org/10.18653/v1/N16-2002

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. ArXiv:1806.03537 [Cs]. http://arxiv.org/abs/1806.03537

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. ArXiv:2007.11464 [Cs]. http://arxiv.org/abs/2007.11464