Multilingual word embeddings and low resources: identifying influence in Antiquity

Marianne Reboul

Multilingual word embeddings and low resources: identifying influence in Antiquity

Marianne Reboul, École Normale Supérieure de Lyon

Background

The computing of semantic influence is a key part to understanding the development and movement of knowledge throughout the world. It is now made possible by the use of powerful computer hardware, and the correlated development of new machine learning techniques. Semantic influence therefore becomes a relationship to be quantified, graphically represented and analyzed: it is now possible to measure how meaning was altered within a language and between languages.

Although new techniques allow researchers to establish semantic proximity between languages using machine learning for aligning and predicting translations, those techniques are data hungry and tend to be less effective on very low resource languages such as ancient languages. Most of the techniques used to develop translation models rely on parallel data training: this is partially due to the fact that automatic translation is built for modern, period independent, usage, such as commercial translations. However, the lack of parallel data for rarer languages and ancient languages has led to research in using monolingual data training for multilingual vector spaces. Recent developments can overcome the scarcity of data using neural nets to produce linear mapping between languages. In this paper, we would like to apply techniques we tested using monolingual training and mapping to create low resource multilingual embeddings without using massive parallel data. We would both use semi-supervised and fully unsupervised techniques.

Isomorphic semantic spaces

Our research relies on an instinctive, yet bold assumption that Ancient European languages are isomorphic, that is to say that we assume that we could show most languages were built in comparable semantic molds, and share, to a certain extent, part of the meaning of their words, or at least share a common semantic conception of their world. The models obtained while training on monolingual data should therefore be comparable. We would firstly explain how, based on the assumption that ancient languages are isomorphic, monolingual semantic spaces can be mapped in one single multilingual space. Secondly, we would demonstrate that semantic spaces trained on different time periods can reveal potential intertextuality and semantic influences between texts. Thirdly, we would apply those techniques to a specific object, that is to say a corpus of Greek and Latin epic poems,

from Homer toVergil, and see how specific terms and aspects of Latin epics tend to be more influenced or disassociate themselves from the Greek canon.

Methods and preliminary results

For our preliminary results, we used two kinds of corpora, although both could be qualified as microcorpora : the training corpus is a compilation of both Homer's *Iliad* and *Odyssey* (Allen's Greek version (Allen, 1908), Estienne's Latin version (Estienne, 1589)), and a test corpus of ten to fifteen texts (TEI format, already lemmatized by Perseus (Smith et al., 2000) or lemmatized in python through pie-extended (Manjavacas et al., 2019)), both in prose and verse, of canonical Greek and Latin epics, such as Vergil's Aeneid or Lucan's *Pharsalia* (although certain texts are not, strictly speaking, epic poems). We also built a very small custom Greek to Latin dictionary, based on two French dictionaries (although we intend to modify this step in future experiments), for the training phase.

If we consider Greek and Latin as isomorphic languages, we would expect some of the terms used in one language to find their most probable correspondance in the other according to their context. That is to say, their semantical environment should be approximately comparable. For example, "Jupiter" and "Ζεύς" should share space with comparable semantic terms used in the same context. If we try to describe this phenomenon with a 3-dimensional semantic space (although our spaces generally have 200 dimensions), one should then see semantic spaces in both languages as two clouds of terms having roughly the same aspect. The method to associate both spaces (in our case trained either with GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013)) would be to move one target space to fit the other. Several methods exist to do so, but the one that gave most effective results on our corpora is the VecMap (Artetxe et al., 2018) method, using linear transformation to map isomorphic spaces. Once the crosslingual semantic space has been created, we can see what vectors one language is more likely to share with the other. However, this is not yet an understanding of "influence", strictly speaking, as it lacks the aspect of evolution through time. This is the third part of the training (after monolingual training and crosslingual training), that is to say historical linking between crosslingual semantic spaces. Our results are not definite at this point yet, but we intend to use dynamic link prediction algorithms (eg. methods described in (Kutuzov et al., 2018)) to measure the evolution process of corsslingual vectors through time.

In the following example (Figure 1), which is a representation of random crosslingual vectors found on two microcorpora (Homer's Greek Iliad and Sommer's French Odyssey), one can see the accuracy of corsslingual associations on very scarce data. Further figures will be included during the presentation (which were not included here due to black and white publication guidelines).

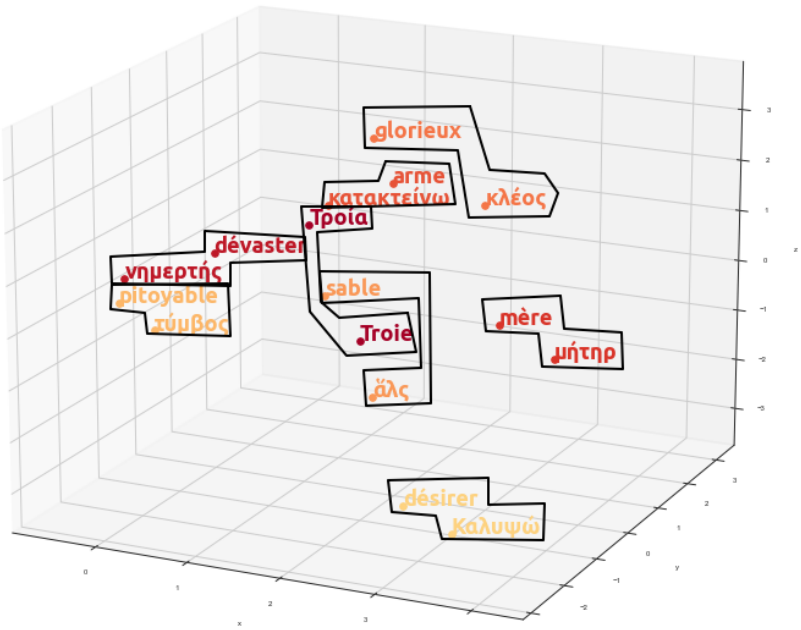

Figure 1 : Preliminary crosslingual semantic space (French-Greek) with nearest neighbours highlighted

Future prospects

This paper would be a first step towards a larger project we intend to develop in the future, which could be summed up in three points : we would identify semantic influences in Ancient European written languages on one another first by building a corpus, in standardized TEI, of at least 50 million words (from 1450 B.C. to 400 A.C) first in Greek, Latin, then extending it to oldest and rarest European written languages (such as Etruscan, Minoan, Iberian, Linear B) ; we then would quantify semantic influences by computing a multilingual semantic space using crosslingual sentence and word embeddings from monolingual pre-training ; lastly, we would show these influences by analyzing and graphically representing chronological correlations between those languages. I therefore hypothesize that I can measure and show the interaction and influence of European Ancient written languages on each other through time.

References

[1]. ALLEN, Thomas William, *et al.* (ed.). *Homeri Ilias*. Clarendon Press, 1908.

[2]. ALLEN, Thomas William, *et al.* (ed.). *Homeri Opera, Tomus III, Odysseae*. Clarendon Press, 1961.

[3]. ESTIENNE, Henri, *Homeri Poemata duo, Ilias et Odyssea, sive Ulyssea*, Genève, Estienne, 1589.

[4]. SMITH, David A., RYDBERG-COX, Jeffrey A., et CRANE, Gregory R. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 2000, vol. 15, no 1, p. 15-25.
[5]. MANJAVACAS, Enrique, KÁDÁR, Ákos, et KESTEMONT, Mike. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*, 2019.
[6]. PENNINGTON, Jeffrey, SOCHER, Richard, et MANNING, Christopher D. Glove: Global vectors for word representation. In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. p. 1532-1543.
[7]. MIKOLOV, Tomas, CHEN, Kai, CORRADO, Greg, et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[8]. ARTETXE, Mikel, LABAKA, Gorka, et AGIRRE, Eneko. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*, 2018.
[9]. KUTUZOV, Andrey, ØVRELID, Lilja, SZYMANSKI, Terrence, et al. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*, 2018.