A case study of diachrony across two languages

Syrielle Montariol[1,2], Alexandre Allauzen[1]

[1] *LIMSI-CNRS, Université Paris-Sud, Université Paris-Saclay,* [2] *Société Générale*

syrielle.montariol@limsi.fr, alexandre.allauzen@limsi.fr

The way words are used changes throughout time. This evolution can be tracked by training time-varying word embeddings on a temporal corpus. Here we extend the analysis of this phenomenon across multiple languages, studying "cross-lingual drifts": the temporal evolution of the representation of the same word in two languages.

We can consider for example the impact of an event on two communities. Its media resonance, represented by a change in the context of involved words, can differ in intensity and form between communities speaking different languages. Detecting these disparities can help understanding disagreements among communities, or evaluating the extend to which some communities are influenced by a given trend or thinking. As a preliminary study, we propose an experimental framework to compare word meaning evolution across two languages using diachronic embeddings alignment of representation spaces.

We rely on two newspaper corpora ranging from 1987 to 2006, divided into 20 yearly time slices: The New York Times Annotated Corpus (NYT) in English, and a corpus of French articles from the newspaper Le Monde. We use the Dynamic Bernoulli Embeddings model (DBE), a temporal version of a probabilistic generalisation of the CBOW model: each word has one embeddings vector per time slice and a unique context vector fixed over time.

To compare the evolution of a given word in both corpora, we first build a bilingual vocabulary by translating and merging the French and English vocabularies from our corpora. Then, we train monolingual word embeddings on each full corpora, normalize it, and align it relying on the bilingual dictionary. Finally, the aligned embedding vectors are used to initialise the dynamic model DBE which is trained separately on both corpora.

For each word, we compare: (a) its drift in the corpus it comes from, (b) the drift of its translation in the other corpus, and (c) the drift of the similarity between the word and its translation. Thus, we differentiate four kinds of cross-lingual drifts (table below): (1) Words that drift in the same direction on both languages; (2) Words that drift on both languages but whose cross-lingual similarity diverges between the first and the last time step; (3) Words that drift in only one language; (4) Words that are stable in both languages.

Classes	1	2	3	4	5
Proportion	5.4	5.5	16.1	15.2	57.8
Example	Renewable	Soviet	Francs	Homeland	Soap

A limit to this approach is the smoothing of the disparities between the two language during the alignment. An improvement could be to use a soft alignment method to decrease its impact on the vectorial spaces.

References: Lample G., Conneau A., Denoyer L. & Ranzato M. (2017). Unsupervised machine translation using monolingual corpora only. arXiv preprint:1711.00043. Rudolph M. & Blei D. (2018). Dynamic embeddings for language evolution. In Proceedings of the 2018 World Wide Web Conference. Tahmasebi N., Borin L. & Jatowt A. (2018). Survey of computational approaches to diachronic conceptual change. CoRR, 1811.06278.