# Measure and Evaluation of Semantic Divergence across Two Language (Syrielle et al.)

Due to the complexity and the difference between each language, which change per time in terms of the meaning of words, Authors in this paper propose to track theses divergences by comparing the evolution of a word and its translation across two language. To achieve their goals, they first investigate on French and English before applying theirs methods to newspaper corpora to detect bilingual semantic divergence and provide qualitative insight for the task. Those methods consist on building time-varying and bilingual word embeddings, using contextualized and non-contextualized embeddings. At the end they evaluated those methods by generating a corpus of synthetic semantic change across the two languages.

The goal of this paper to extend the analysis of lexical semantic change across two languages, aiming at estimating the degree of diachronic semantic divergence between a word and its translation across time in a bilingual corpus.

What I like about the paper is that:

- They presented a system based on contextualized embeddings which is an important fact while working on semantic of words
- Instead of training the model from scratch, the fine-tuned the model at each time step
- For the Diachronic training, the used two methods: incremental training and independent training .
- They employed a form of supervision to obtain cross-lingual word embeddings (Anchoring) used as seeds during the alignment which will force the embeddings of the word pairs from the supervision dictionary to be the same in the two languages.
- For the monolingual setting, the used two metrics commonly used to measure the drifts of a word in each language: the incremental drift and the inceptive drift.
- To generate and capture variations of distributions of word senses through time and across two languages, they defined several scenarios of word usage variations.
- They used the drifts obtained from the measures of sense distribution as a gold standard for the valuation of their systems.
- The best results are obtained with BERT using k-means clustering over the others scenarios
- Words that are stable in both languages (B0) are mostly daily life words (e.g. mayonnaise). Words that drift in the same direction in both languages (B2) are concepts related to technology and society that are common to the English and French culture (e.g. renewable); while the words that diverge between the two languages (B1-fr (English stable, French drifting), B1-en and B3) belong to more culture-specific concepts (e.g. francs) or controversial topics (e.g. terrorist).

  **Deficit:** lack of bilingual datasets annotated with semantic divergence, getting a high quality of diachronic bilingual representation, they use of an injection to define word pairs, the evaluation with synthetic data

What I suggest to improve in this paper is to also measure and classify the semantic convergence and be investigate in the quality of diachronic bilingual representation, the metric to measure the semantic divergence and finally explore others options of detecting semantic divergence.

# BERTScore: EVALUATING TEXT GENERATION WITH BERT

As commonly used methods rely on surface-from only, simply counts n-gram overlap between the candidate and the reference and fails to account the meaning-preserving lexical and compositional diversity, in this paper, authors propose BERTScore, an automatic evaluation metric for text generation based on pre-trained BERT contextual embeddings. The metric computes a similarity score for each token in the candidate sentence with each token in the reference sentence. They demonstrated that BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics and finally showed that BERTScore is more robust to challenging examples when compared to existing metrics.
The goal is to evaluate semantic equivalence.

What I like about the paper is that:

- They used a contextual embeddings, which captured the specific use of a token in a sentence, and potentially captured sequence information.
- Their approach was relatively simple and portable to new languages as they did not used external tools to generate linguistic structures.
- BERTScore is not optimized for any specific evaluation task.
- BERTScore enables them to easily incorporate weighting using idf scores computed from test corpus.
- The robustness was tested using adversarial paraphrase classification
  Deficits: there is no one configuration of BERTScore that clearly outperforms all others metrics

What I can suggest to improve the methods in this paper is to apply the BERTScore experiments on others NLG tasks such as Chatbots or Questions and Answering to see how it can make improvements.