Anchor points for bilingual lexicon extraction from small comparable corpora

Bilingual Lexicon Extraction

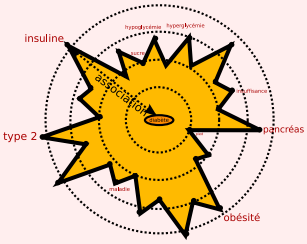
Identification of translations among comparable corpora

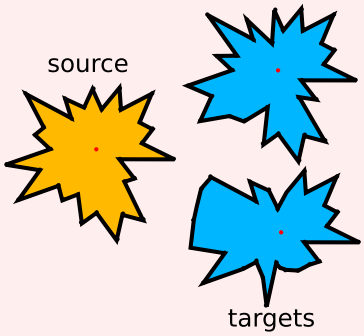
- Harvest surrounding words around a given term
- Compute association score for each surrounding term

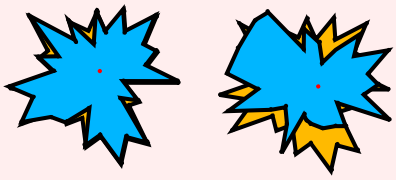
➡ Association pattern for each term, in source and target language

- Compare source and target patterns

Similar patterns ➡ translations candidates







Anchor points

- Automatically identifiable
- Relevant, regarding corpora topics
- Not ambiguous (no polysemy)

➡ Use to improve discriminative power of context vectors

Anchors points used

Japanese transliterations

- written using a dedicated set of symbols
- can be bound to French or English based on pronunciation
- in scientific context, reflect terminology of the domain

Scientific compounds

- can be detected with a list of affixes on French and English
- bound to Japanese using dictionaries
- frequently used in scientific contexts

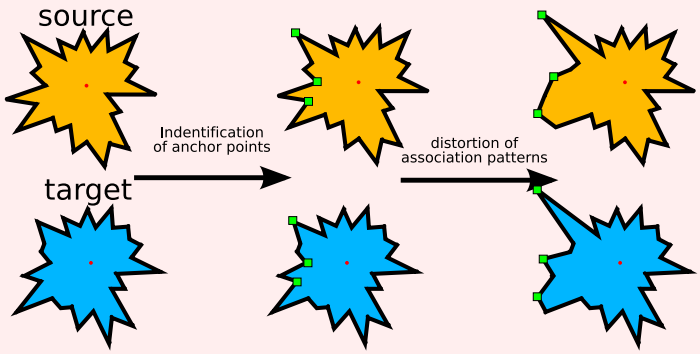
Results, conclusion

- Improvement of base results for English/Japanese and French/Japanese alignment, especially when using scientifi compounds
- Moreover, massive re-ranking of translation candidates

➡ Method easy to use and implement, and to adapt to different languages

- Relevant for small comparable corpora, where there are not enough occurrences of a given term to have a fine description of that term

Idea: using anchor points to distort association patterns



Emmanuel Prochasson[1], Emmanuel Morin[1] and Kyo Kageura[2]

[1]Laboratoire d'Informatique de Nantes-Atlantique
Université de Nantes

[2]Graduate School of Education
University of Tokyo