Distributional Semantics

Lecture 9. Cross-lingual Embeddings and Machine Translation

Florian Gouret

April 20th, 2019

Before hands:

This seminar is mainly based on the work of S.RUDER. More information about cross-lingual word embeddings can be found here:

http://ruder.io/cross-lingual-embeddings/index.html

Presentation of the seminar

Plan:

1) Introduction and history
2) Different supports and embeddings
3) Conclusion and practical

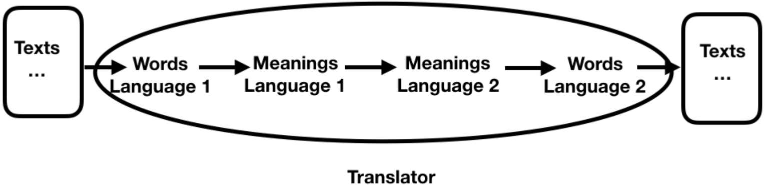
Introduction and history: A science not that old

Any thoughts about machine translation?

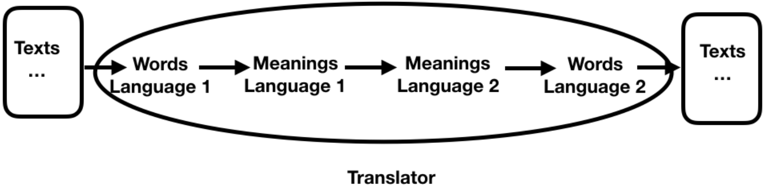
Introduction and history: History

1629, René Descartes proposes a universal language
1949, creation of the field "Machine-translation"
1951, Yehosha Bar-Hillel, MIT
1970, French Textile Institute, French, English, German and Spanish
1971, Brigham Young University, automated translation

Introduction and history: Process of translation

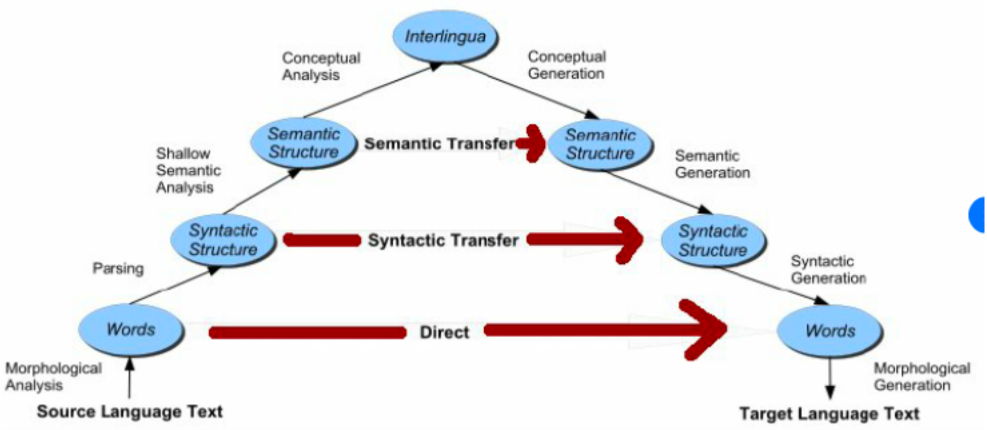


Introduction and history: Process of translation



But how to really translate? What are "meanings"?

Introduction and history: Vauquois' triangle



Different supports: Presentation

According to tasks, the amount of data and the quality of them can different. For instance, translating all words from a dictionary without their definitions can be done more easily than translating automatically the last book or news paper. Hence **three main levels** can be distinguished:

- word alignment
- sentence alignment
- document alignment

According to each level, the quality of translation may differ; the more complex sources are, the more difficult the translation is going to be. Nowadays, it is possible to say that for some simple tasks, computers can automatized almost perfectly translation processes however, going to daily life texts is no perfect yet.

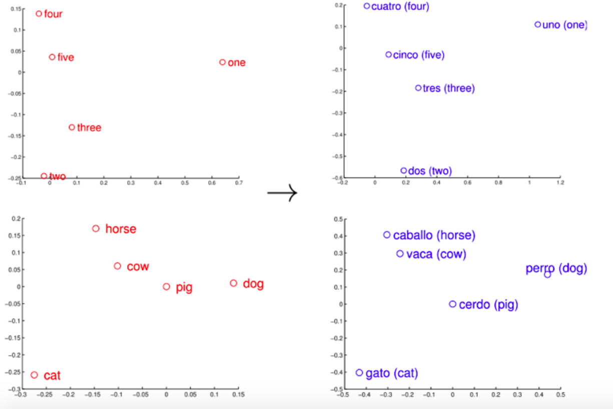
Different supports: Word alignment

Given two languages, the word level alignment tries to find the best way to link words with the same meaning. For instance aligning the English words "I "saw "a"and "cat"with the French one "J "ai "vu "un"and "chat in the sentence "I saw a cat "J’ai vu un chat". As it may appear, obtaining these dataset is an expensive task as the accuracy is really big.
Different variants of word alignment have been explored such as:

- Mapping based approaches
- Word level based on pseudo Bilingual corpus
- Joint models
- Word level alignment methods with comparable data

Mapping based approaches: Minimizing error

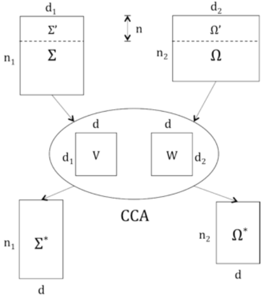
Using the 5000 most frequent words of two languages, a mapping between each representation tries to be found.



The function which is minimized a stochastic gradient $\Omega_{MSE} = \sum_{i=1}^{n} \| WX^S - X^T \|^2$

Mapping based approaches: CCA based mapping

The canonical correlation analysis takes two languages as input and tries to find correlation between them.



Through this method found by Faruqui and Dyer, antonyms and synonyms are separated which was not the case with the previous method.

Sentence alignment:

The sentence alignment focuses on collection of words instead of all same words. Hence obtaining data is a bit less complex. They may be collected from talks or speeches on live translated or from translated books.

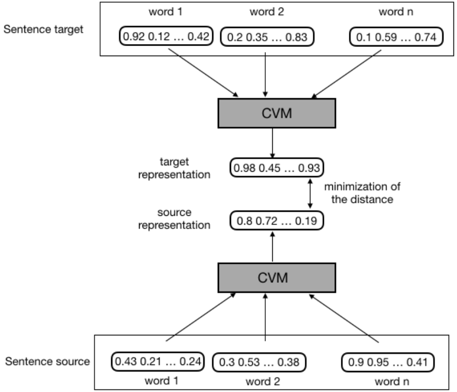
Here are some methods dealing with sentence alignment:

- Bilingual compositional sentence model
- Bilingual auto-encoder
- Bilingual skip-gram

Sentence alignment: Bilingual compositional sentence models

The bilingual compositional sentence model was developed by Hermann and Blunsom. The model trained is based on two models. These two give a representation of aligned sentences and try to minimize the distance between them:

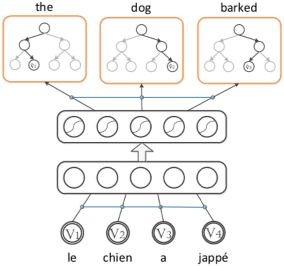
$$E_{dist}(R_T, R_S) = \| R_T - R_S \|^2$$



Sentence alignment: Bilingual auto-encoder

This model, by Lauly et al., is based on a learning process which aims to reconstruct the bag-of-words representations of aligned sentences. The main advantage of this model is its capacity which outperformed some other algorithms without using word level alignment. Based on errors and updates, it reconstructs sparse binary vector of word occurrences. Train autoencoder and decoder using softmax to reconstruct sentences and translations.

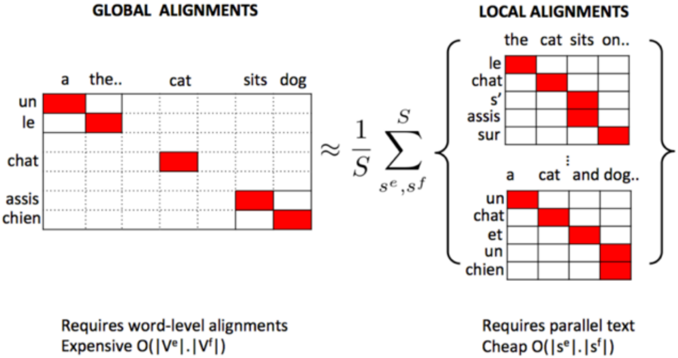
$$J = L_{AUTO}^{S\to S} + L_{AUTO}^{T\to T} + L_{AUTO}^{T\to S} + L_{AUTO}^{S\to T}$$



Joint level: Bilingual bag-of-words

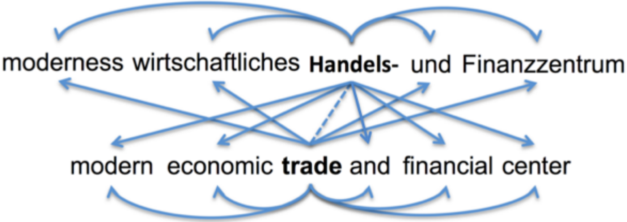
With this model, the assumption made by Gouws et al. is that each word in a source sentence is aligned with every word in the target sentence under a uniform alignment model.

Hence, the minimization does not focus on the distance between aligned words but minimizes the distance between the mean of the word representations in the aligned sentences.



Joint level: Bilingual skip-gram

The skip-gram model is based on the context of occurrence of each word. From this, a representation is given. Luong et al. extended to cross-lingual. They used words from the source to predict their aligned in the target language.


moderness wirtschaftliches **Handels-** und Finanzzentrum
modern economic **trade** and financial center

Document level alignment: Pseudo bilingual document aligned approaches

The method proposed by Vulić and Moens is not to merge monolingual corpora into one but to merge two aligned documents into a pseudo bilingual document. Once concatenated, a shuffle is done permuting words.
A major issue with this method is that results can be as bad as good, also the random permutation was modify into a more linear one: words can be insert in the order they appear in the original source.

Document level alignment: Concept based models

Søgaard et al. in their model decided to consider that articles of Wikipedia across languages share the same concept and then it could be possible to learn a representation not concept based but word based. Indeed, pages of a same topic within different languages usually use words in the same way.

Evaluation:

For each model, an evaluation of performances has to be conducted. Different variants are possible with for all advantages and drawbacks.

- Word similarities
- Bilingual dictionary induction
- Benchmarks

Evaluation: Word similarities

The question is to see how results match with human decisions. Hence large lists of paired words, source language and target language, are made and then checked with results.
With SemEval 2017, a cross similarity dataset was introduced.
One of the main drawbacks of the metric is that it can not handle polysemy and thus can provide unappropriated results even this a good score.

Evaluation: Bilingual dictionary induction

This evaluation is freely available and can be used to evaluate performances on not common language.
Dictionaries are hand made and thus rather accurate even if some issues with polysemy can occur.
For specialized texts, results are very good if the spelling is taken into account.

Evaluation: Benchmarks

Some websites offer to evaluate word representation obtained as such as http://wordvectors.org or http://128.2.220.95. If both offer an easy to use solution, kernels are different. The first one focuses mainly on evaluating the word representation based on a word similarity dataset whereas the second one can evaluate more results such as word similarity, multiQVEC, bilingual induction, document classification.

Questions?

Based on the code given by MUSE, build a simple translator between two close languages (for instance Russian-Ukrainian/Belorussian, English-French, Swedish-Norwegian). To do so, we assume that the translation of a given word in a targeted language is the closest neighbor once the source basis is aligned.

Resources:

- https://github.com/facebookresearch/MUSE/blob/master/demo.ipynb
- http://ruder.io
- https://fasttext.cc