It's all Greek to me! learning to translate languages

Semester or Master project

Keywords: *big data, machine learning, machine translation, natural language processing*
Supervisors: Florent Garcin, {*firstname.lastname*}*@epfl.ch*

Description

Machine translation uses software to translate text from one language to another. One naive approach is to perform a simple substitution of words in one language to another. However, this approach alone does not produce good translation because it does not take into account the whole sentence.

Techniques using corpus and statistical models are getting popular and bring better translations. Usually, a statistical machine translation algorithm is decomposed into three components [1]: translation model, language model, and decoder. Translation models could be based on words, phrase or syntax. These models transform a source phrase into a target phrase. Language models indicate how good a sentence is in the target language. It is used as a verification step after the translation model to resolve ambiguity. Finally, the decoder search for translation with the highest score based on to the models learnt before.

We are interested here in a semi-supervised learning approach in which we will be able to train our algorithms on partially-labelled data. The goal of this project is to survey the state-of-the-art algorithms in machine translation, implement and evaluate several algorithms, and possibly improve them.

Tasks

1. Survey the state-of-the-art algorithms in machine translation
2. implement the algorithms
3. evaluate the algorithms
4. improve existing approaches
5. optional: take part in one of the international competitions on machine translation

References

[1] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.