DISAMBIGUATION OF JAPANESE ASPECT-MARKERS WITH HELP OF PARALLEL-CORPORA

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

29.05.2017

Rule-based systems are a ”classical approach” in machine translation, the linguistic infroamtion is retrieved from dictionaries and grammars of each language respectively. Such systems are unconvertible for a lot of methods in Computerlinguistic, because the rules are too abstract. Therefore, example-based system are easier to convert if the data is available. Taking into account the enormous difference between English and Japanese, such type of translation is a nice solution. This was first suggested by Makoto Nagao.

There are different types of parallel corpora that are used for this project, for instance:

- Wikipedia Corpus
- Basic-Sentences Corpus
- ”Wachturm”

In Japanese aspect-markers, depending on the context, may deliver different aspects. In English the ”running form” is built through Progressive, however, a state can not be expressed in this form.

First of all, we should prepare the data by creating partial corpora by filtering all the sentences that do not receive a certain construction. Secondly, align the verbs and thirdly, parse and determine the tense of the English verb. Hand-annotated data is usually very effortful to create. Parallel corpora, on the other hand, are pretty common and are easily accessible. In case of translations, ambiguous constructions must be disambiguated by the translator.

Classifier:

- Introducing data in two different types: training and test
- Using different algorithms for classification

The goal is to train the classifier in order to disambiguate aspect-markers in Japanese. The categories of the training data should not be annotated, but taken from the respective translation. A problem worth mentioning is that the Japanese aspect-marker is not congruent with the English tenses. Evaluation of the achieved accuracy is done by using test data. Using some alignment models of GIZA++ (an extension of the program GIZA, a part of the Statistical Machine Translation) deliver quite poor results, about 30 per cent only.