Improving speech translation by automatic punctuation of speech recognition output

2015–2016


In the context of a project on Computer-Aided Translation (CAT) we are working on the fascinating task of speech translation. This means that we have an audio signal containing speech in a source language text (in our case Dutch) and we want to acquire the corresponding text in a target language (here English). Traditionally this is attempted by using the output of an automatic speech recognizer (ASR) as input for a machine translation system. This cascaded method of tackling the problem however has several disadvantages. First of all, it has been shown that MT systems perform a lot worse when they are confronted with the imperfect output of an ASR system which potentially contains many recognition errors. Moreover, the ASR output does not contain punctuation which is valuable information for the MT system. Finally, separate optimization of the 2 systems (typically on different data, using different evaluation criteria, ...) is bound to be worse than joint optimization.

In this thesis/internship we want to focus on correcting the output of the ASR system. You will be given Dutch transcriptions and a Dutch-to-English MT system and your task then is to add punctuation to the transcriptions such that the final translation is more accurate. This can be done in different ways: one possibility is to use a language model which predicts the next word (or punctuation mark) in a sentence, given the previous words. Another possibility is to identify phrases using a (shallow) syntactic parser and apply punctuation based on rules. Finally, a general machine learning approach can employed which can use arbitrary features to predict punctuation.

Programming skills are required and experience with one or more of the following is beneficial:
- statistical language models
- syntactic parsers
- machine learning

Promoter
Prof. Patrick Wambacq (01.04)

Supervision
Joris Pelemans (02.86), Lyan Verwimp (02.86)

Workload
- literature and study - 25 %,
- implementation - 25 %,
- experiments - 50 %

Number of students
1