

2007 Student Research Symposium

Language Technologies Institute



Noun Phrase Driven Bootstrapping of Word Alignment

Vamshi Ambati

Abstract:

The work described here is within the AVENUE project at LTI, that aims at building Machine Translation systems for resource poor languages. AVENUE primarily concentrates on language pairs where one language is a resource-rich language like English and the other is a resource poor language like 'Mapadungun'. The strength of AVENUE lies in its ability to combine the intuitive and more natural linguistic phenomenon with the statistical and empirical evidence obtained from large data. The system uses grammatical transfer rules to represent the translation process between two languages and a statistical decoder to finally select the appropriate target translation from a set of hypothesis created during the translation process.

My work in this project is particularly related to learning the grammatical transfer rules automatically from data.

My work in AVENUE project is particularly related to automatically learning the grammatical rules of translation from large unsupervised data. One of the languages in the translation is assumed to be resource rich, consisting of parsers, language annotators and other tools to analyze the language. Therefore, the task of rule learning now boils down to being able to capture as rules, the process of how language concepts in one language manifest or project themselves into the other language. We will call this 'Syntax Projection'. As we are provided with only unsupervised data, one basic idea to perform syntax projection is, to automatically word align the corpus and heuristically transfer the syntax via the word alignment.

Although this approach seems promising and has been applied in recent years, two main problems have to be tackled. Firstly language exhibit divergences and so linguistic concepts may not transfer directly from language to another. Secondly the word alignment that is used as a bridge to transfer syntax is not reliable. Although many automatic Word alignment models have been proposed and applied with great success, they still do not provide high quality alignment when less data is given for training and also when the languages are diverse. This is in fact the case in the AVENUE project, as we deal with language pairs that are quite diverse and where large resources can not be assumed.

Therefore this work discussed here addresses the particular problem of being able to reliably project syntax, in particular noun phrases when word alignment is noisy.

We mentioned in the previous section, the two main problems - divergences in language pairs and the word alignment quality - that prevent the successful projection of syntax which could then be used in rule learning. We take an approach that tries to address both these problems.

We address the first of the problems by concentrating only on the projection of Base Noun Phrases (base NPs). Noun phrases do not diverge from one language to the other as they represent 'semantic concepts' that are also syntactically consistent and contiguous. We then propose an iterative noun-phrase driven bootstrapping approach to simultaneously improve the projection quality and the word alignment quality. Our approach is to use the initial automatic word alignment to extract all the Noun Phrases and their syntactic projections via this word alignment. We then use some heuristics and some statistical techniques to reliably prune the huge list and obtain a clean and more reliable NP table after projection.

Using the NP table thus obtained, we then use the technique of 'constrained word alignment' to automatically realign the sentence pairs. Constrained word alignment technique is basically to only align words within a source side of the noun phrase with words on the target side of the noun phrase. Similarly words outside of the source noun phrase boundaries are aligned with those outside the target noun phrase boundaries.

The above process is iteratively performed to obtain cleaner word alignments and as a result also obtain cleaner Noun Phrase projections.

Since two tasks are simultaneously tied into an innovative bootstrapping technique, we try to evaluate the improvement in both these tasks.

Our evaluation of obtaining a clean NP phrase projection is verified by checking against a human created NP phrase table for a given set of parallel sentences. The evaluation metric we use here is Precision, Recall, F scores of the NP phrases extracted after projection vs. the human extracted gold standard NP table.

Similarly we also verify the word alignment quality obtained by using the bootstrapping, constrained alignment approach. For this we train the system on 55K sentences and test it on 200 sentences for which gold standard human alignment is present. Standard metrics of Alignment Error Rate (AER), Precision, Recall and F-measure are used to evaluate the word alignment task.

Initial results have shown great promise and we could observe over iterations that as a result of the bootstrapping over subsequent iterations, one can find improvement in the Precision and Recall of the NP projection and also a decrease in the Alignment Error Rate of the overall word alignment. The NP projection increases from 40% F1 to 41% in just one iteration where as the AER reduces from 55 points to 54 points.