# Towards Syntactically-Contrained Statistical Word Alignment

Greg Hanneman
11-731: Advanced Machine Translation Seminar
Project Proposal

January 23, 2008

## 1 Scope and Approach

The goal of this project is to review current models and techniques for automatic statistical word alignment, with a particular view towards considering the problem of word alignment when also taking syntactic or other linguistic information into account.

We will begin with an overview of the basic IBM statistical models and the foundation of the modern notion of statistical word alignment. We will also review any competing models or heuristic approaches that have found success on alignment tasks or in real machine translation systems.

GIZA++, a popular (and perhaps the most used) toolkit for parallel corpus word alignment, will also be discussed, particularly as an implementation of the IBM and HMM models. We will pay special attention to any support it offers for constraining the alignment problem in ways similar to those described below.

The main current and future research direction this project will address is the question of how word alignment can be done under the constraint of any syntactic or other linguistic knowledge that might be available for the parallel corpora being aligned. For example, if part-of-speech tagging and noun-phrase bracketing were applied to the parallel SVO sentences

```
[les étudiants] vont à [l' école]
[the students] go to [school]
```

we would like a way to ensure that any alignments for words within the NP "`the students`" fall within the scope of the corresponding NP "`les étudiants`." This project will therefore conclude by examining techniques or models that already allow this type of linguistic information to play a role in alignment, and will also discuss what might be required to extend other existing approaches to be able to do so.

At the conclusion of the project, we hope to have been able to map out the current state of the art in applying syntactic constraints to statistical word alignment, and to suggest concrete ways in which a viable toolkit to do alignment under those constraints might be built and applied within an MT system.

## 2 Initial Readings

- Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer (1993). "The Mathematics of Statistical Machine Translation." *Computational Linguistics*, 19(2).
- Franz Josef Och and Hermann Ney (2003), "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, 29(1), 19–51.
- John DeNero and Dan Klein (2007). "Tailoring Word Alignments to Syntactic Machine Translation." *Proceedings of ACL 2007*, 17–24.