ABSTRACT

Title: Statistical Machine Translation between Languages with Significant Word Order Differences
Author: Bushra Jawaid
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Daniel Zeman, PhD.
Supervisor's email address: zeman@ufal.mff.cuni.cz

Abstract:
One of the difficulties statistical machine translation (SMT) systems face are differences in word order. When translating from a language with rather fixed SVO word order, such as English, to a language where the preferred word order is dramatically different (such as the SOV order of Urdu, Hindi, Korean, ...), the system has to learn long-distance reordering of the words. Higher degree of freedom of the word order of the target language is usually accompanied by higher morphological diversity, i.e. word affixes have to be generated based on the fixed word order in the source sentence.

The goal of the thesis is to explore the two mentioned (and possibly other related) classes of problems in practice, and to implement and evaluate techniques expected to help the SMT system to solve them. This includes:

1. Selecting a language pair with word order differences and collecting parallel data for the pair.

2. Training an existing SMT system on the data.

3. Evaluating the performance of the system and analyzing the errors it does. Estimating how much the accuracy of translation is affected by the problems mentioned above, and possibly what are the other types of error causes that dominate the output.

4. Implementing preprocessing and/or other techniques aimed at minimizing the found classes of errors. Evaluating their impact.

Keywords: Statistical Machine Translation, syntactic word order differences, rich morphological languages, parallel corpus