Vincent Vandeghinste

A Hybrid Modular Machine Translation System

In this thesis, we present the LoRe-MT system (Low Resources Machine Translation) to translate Dutch to English. It is a demonstration of techniques for constructing MT systems for languages lacking extensive resources like bilingual corpora or complex parsers (so called low resource languages). We develop and test methods for relatively cheaply and easily building MT systems for new language pairs where extensive pre-existing resources may not be available.

This thesis describes the system in detail, the resources used in its development, and extensively evaluates it and its components.

The LoRe-MT system is a response to several important MT implementation problems:
- How far can we go in implementing quality MT systems using only basic tools and rudimentary resources like shallow source language analysis, bilingual dictionaries and processed monolingual target language corpora?
- Can we resolve word and phrase order issues using exclusively target language corpora?
- Is it possible to perform adequate lexical selection (choosing the correct translation from a set of candidates) using only target language corpora?

This thesis includes a general introduction to the problems posed by machine translation, and proceeds to describe in detail the processes used in translation:
- Minimal resource shallow source analysis
- Source to target language transfer using a bilingual dictionary and a very small rule set
- Matching outputs to the contents of target language corpora in order to extract the most likely correct target language sentence structure

This system and the implementation choices we made are evaluated in depth, and we discuss the potential for each component to be reused in other natural language processing tasks. We also provide an outline for future work and development of the LoRe-MT paradigm.

Keywords: hybrid machine translation, low resource languages, Dutch to English translation