Title: Hybrid Machine Translation Approaches for Low-Resource Languages

Autor: Amir Kamran

Department: Intelligent Computer Systems
Faculty of Information and Communication Technology
University of Malta

Supervisor(s): Professor Michael Rosner & Mgr. Martin Popel

Abstract: In recent years, corpus based machine translation systems produce significant results for a number of language pairs. However, for low-resource languages like Urdu the purely statistical or purely example based methods are not performing well. On the other hand, the rule-based approaches require a huge amount of time and resources for the development of rules, which makes it difficult in most scenarios. Hybrid machine translation systems might be one of the solutions to overcome these problems, where we can combine the best of different approaches to achieve quality translation.

The goal of this dissertation is to explore different combinations of rule based and semi automatic preprocessing techniques for English-to-Urdu statistical machine translation and to evaluate their performance over the standard corpus based methods currently in use. This includes:

1. Insertion of artificial linguistic markers in English to improve the word alignment from English-to-Urdu.

2. Use of syntax-based word reordering rules to tackle the long distance reordering problem in statistical machine translation.

The novel element in the proposed work is to develop an algorithm to learn automatic reordering rules for English-to-Urdu statistical machine translation. Moreover, a comparison between hand written reordering rules with automatically learned rules will also be a part of this dissertation.

Keywords: machine translation, long-distance reordering, automatic rules extraction, language markers, low-resource languages, source-side preprocessing