Short Term Scientific Mission
Uppsala University
Department of Linguistics and Philology

Host: Prof. Joakim Nivre
Visitor: Meghdad Farahmand

December 22, 2014

Abstract

During this visit which lasted three months, we studied two divergent approaches to extraction of Multiword Expressions (MWEs) based on their internal and dependency-based contextual properties. We further planned and started the creation of an English compound dataset.

1 Studying Extraction Techniques

We studied two main types of idiosyncrasy in MWEs, i.e., statistical idiosyncrasy and semantic idiosyncrasy or non-compositionality. In order to learn statistical idiosyncrasy we modeled non-substitutability of MWEs by looking at the probability of their occurrence and the probability of the occurrence of their constituents' near synonyms. This study resulted in a paper which was submitted to NAACL 2015.

In order to learn non-compositionality, we took a context based approach. For context to be more meaningful, we looked at the parse tree of the sentences which contain MWEs and extracted a grammatically meaningful context vector for each MWE instance. We developed several context based baselines based on intuitive characterization of the context regardless of any linguistic information. We trained and tested linear and quadratic classifiers for different models to see their predictive abilities on non-compositionality. This is a work in progress. It will be studied for 2-3 languages and the results are planned to be submitted to MWE 2015.

2 Dataset Creation

Absence of a MWE dataset that incorporates instances of both MWE and non-MWE classes is notorious in the field. To the best of our knowledge, almost all available datasets list a particular category of MWEs which can be used to evaluate extraction and identification systems in terms of true positive and false negative. True negative and false positive however, can not be acquired from available datasets (i.e. only positive instances). Nevertheless, annotating MWE and non-MWE in a binary way is a hard task. It requires careful planning and multiple discussions. We started the creation of such dataset. The progress has been considerable. This freely available dataset is planned to be presented at MWE 2015.

3 Confirmation by the host institution of the successful execution of the STSM

On behalf of the institution, I can confirm that Meghdad Farahmand's STSM at Uppsala University has been very successful. Besides making significant progress on his MWE research, as described above, Meghdad has contributed greatly to the computational linguistics group at Uppsala University through his participation in seminars, courses and other activities.

Joakim Nivre, Professor of Computational Linguistics, Uppsala University