THESIS PROPOSAL – 30 CREDITS

Machine Learning for Named Entity Recognition and Classification in Swedish Text

Background

Named Entity Recognition and Classification (NERC) is a generic task in natural language processing. While a given application may have special requirements as regards the entities that should be recognized, some categories tend to show up quite often, among these People, Organisations and Locations.

NERC systems have been developed also for Swedish, but they have often been domain-specific and are hard to compare as they have been evaluated on different data sets.

Project description

An ongoing project within the Swe-Clarin infrastructure is creating new data for testing and training NERC-systems with a focus on eight different categories. The purpose of this project is to develop a Swedish NERC-system for these eight categories using machine learning methods, including neural ones, the new data set, and whatever other data can be found.

A goal of the project is to estimate what levels of precision and recall that can be achieved for the eight categories: People, Organisations, Locations, Events, Products/Works_of_Art, Times, Medical problems, and Medical treatments.

New methods for NERC-systems are proposed and evaluated at regular intervals for many languages such as English, German or Spanish. These methods should be studied as part of the work, and an informed choice on method be made for the Swedish system.

Contact at NLPLAB

Lars Ahrenberg, lars.ahrenberg@liu.se

Student profile

Masters student in computer science, cognitive science or statistics with course points in and interest for machine learning, text mining or language technology.