Capstone Proposal:
Complex Word Identification

1. Domain Background

Complex Word Identification (CWI) is considered to be a subtask of Automatic Text Simplification, which nowadays is a widely researched topic in the field of Natural Language Processing (NLP) (Siddharthan, 2014, p. 2).

CWI can be described as a task of automatically translating a text that for various reasons is considered to be complex into its simpler form. A widely used example to illustrate CWI is Simple English Wikipedia, where Wikipedia articles are simplified in order to be easier understood by all readers. However, the ways in which a text should be simplified largely depends on the target reader. For example, people with language impairments such as dyslexia would benefit from different simplifications of texts than people who suffer from aphasia. Language learners or children would require yet again different simplification rules. Such rules are numerous and include lexical, syntactic simplifications, making long sentences shorter or even summarising the original text. However, the first step in such a system would be to identify its complex parts. CWI is a part of this first step since its goal is to identify which lexical units, or words, of a text are considered to be complex.

As in many other tasks in NLP, ATS and CWI in particular are quite well researched when it comes to English but it still requires more attention when it comes to other languages.

2. Problem Statement

In my project, I would like to build a CWI system for Sweedish in this way contributing to CWI research in other languages than English. Because of the growing mobility in the world and as a language learner myself, I would also like to select second language learners as the target group on which the complexity of the words will be determined. The problem that I would like to solve in my project is therefore automatically classifying Swedish words into classes that represent their lexical complexity corresponding to the language proficiency needed in order to use or understand such words.

3. Datasets and Inputs

I will use Complex Word Identification dataset for Swedish that consists of over 4000 Swedish words annotated with 4 complexity levels (Smolenska, 2016). These complexity levels are annotated based on the needed language proficiency level of the user of such words:
Level 1 words can be used and understood by the beginners, level 2 words - intermediate learners, level 3 - advanced learners and level 4 - fluent speakers of the language.

4. Solution Statement

The solution to the problem will be largely based on the SemEval task of Complex Word Identification system for English (Paetzold and Specia, 2016).
The features used for training the system will be based on word form (such as length or number of vowels) and their frequency in a corpus. The baseline system will be a model trained using only 1 or 2 of these features. Since research shows that simple statistical models often reach satisfactory results in this task, the tested models will be SVM classifier and RandomForrest classifier. If needed, more sophisticated models will be considered.

5. Benchmark Model

The benchmark model will be SVM classifier trained on two linguistic features: word length and its frequency in a corpus.

6. Evaluation Metrics

Accuracy will be used to measure the performance of the models.

7. Project Design

- Building a dataset: adding features to the CWI for Swedish dataset
- Splitting the system into training and testing sets
- Training and testing SVM and Random Forest classifiers. If the results are below 70%, testing more complex models.

The system will be built in Python, using NLTK library and scikit-learn libraries for preprocessing and training.

References

Advaith Siddharthan. A survey of research on text simplification. International Journal of Applied Linguistics, 165(2):259–298, 2014.

Smolenska Greta. Complex Word Identification for Swedish. 2018, Uppsala University.

Gustavo Paetzold and Lucia Specia. Semeval 2016 task 11: Complex word identification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 560–569, 2016.