Introducing SAPIS – an API Service for Text Analysis and Simplification

Daniel Fahlborg, Evelina Rennes

Department of Computer and Information Science, Linköping University, Linköping, Sweden
SICS East Swedish ICT AB, Linköping, Sweden
danfa407@student.liu.se, evelina.rennes@liu.se

Abstract

In several projects, we are developing tools and techniques for simplifying and analyzing textual data, aiming to enhance the accessibility of texts. We present SAPIS, an API service by which these techniques can be reached from a remote server. The API currently involves four running services, and is designed for easy implementation of new services. SAPIS aims to reach professional or daily users interested in the simplification and analysis of texts.

1. Introduction

Within several research projects, we have developed tools and techniques aiming to facilitate the process of creating texts that are easy to understand. Now, we introduce SAPIS, an API service providing the ability to distribute the tools and techniques to a range of professional or daily users. SAPIS currently supports four services of text analysis and simplification, but will be extended to provide the most recent tools and services available. SAPIS is aimed primarily at web editors, but is useful for anyone with an interest in understanding the readability of texts and how they can be simplified.

2. SAPIS

SAPIS is a restful web service based on Java Spring[1], implementing an API with the ability to interpret options and input data as variables in an input JSON object, where the options variable specifies program arguments on a string format. The input JSON object is passed to the SAPIS service in a HTTP request. The open access of the API allows for distribution of any web service uploaded to the same server domain.

2.1 Current services

SAPIS is currently running four services; STILLETT, LEXICALMETRICS, SURFACEMETRICS, and STRUCTURALMETRICS. A client can specify which of these four services to run, together with their respective arguments, by passing instructions with the options variable in the input JSON object. The resulting metrics for text analysis and simplification suggestions are merged and returned to the client as one JSON object.

- STILLETT (Rennes and Jönsson, 2015) is a rule-based automatic text simplification tool for Swedish, partly built on COGFLUX (Rybing et al., 2010), with a dynamic structure of processes and modules, where each process runs a number of modules, but with significant improvement regarding resources and functionality. The preprocessor runs Stagger (Östling, 2013) and MaltParser (Nivre et al., 2007). The tool currently supports rewriting to passive-to-active, quotation inversion, rearranging to straight word order, sentence split, and synonym replacement, in addition to the original rule sets proposed by Decker (2003). The synonym replacement module implemented in STILLETT was originally developed by Abrahamsson (2011).

  When implemented in SAPIS, STILLETT was extended with the option of presenting feedback on a sentence level, based on the same set of rules, but without the actual execution.

- SCREAM (Swedish Compound REAdability Metric) (Sjöholm, 2012) provides statistics about Swedish texts pertaining to their readability. These metrics can be used to provide an analysis of a text at a document level. SCREAM consists of the following services for calculating readability metrics (cf. Falkenjack et al. (2013) for a further description of the feature classifications):

1. The SURFACEMETRICS service provides the main features traditionally used for simple readability metrics. These shallow features utilize tokenization in order to extract features such as *LIX*, *OVIX*, *Nominal ratio*, *Average sentence length* and *Average word length*.

2. The LEXICALMETRICS service provides a categorized frequency analysis from word occurrences in the basic Swedish vocabulary SweVoc dictionary. The word frequencies are extracted after lemmatization, and a high ratio of SweVoc words indicates a high ratio of commonly used words, hence an easy-to-read text.

3. The STRUCTURALMETRICS provides the syntactic and morpho-syntactic features described in Falkenjack et al. (2013). The service provides analysis and metrics consisting of statistical features based on part-of-speech tags and dependency tags.

3. Concluding Remarks

SAPIS provides the ability to distribute tools and services for text analysis and simplification to a range of professional or daily users with the ambition of producing easy-to-read texts. SAPIS currently supports four services of text analysis and simplification, but will be extended to provide the most recent tools and research projects available.

[1] https://spring.io/

References

Peder Abrahamsson. 2011. Mer lättläst – påbyggnad av ett automatiskt omskrivningsverktyg. Bachelor's thesis, Linköping University.

Anna Decker. 2003. Towards automatic grammatical simplification of swedish text. Master's thesis, Stockholm University.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, NEALT Proceedings Series 16.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Evelina Rennes and Arne Jönsson. 2015. A tool for automatic simplification of swedish texts,. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa-2015), Vilnius, Lithuania,.*

Jonas Rybing, Christian Smith, and Annika Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of Texts. Swedish Language Technology Conference, SLTC, Linköping, Sweden.

Johan Sjöholm. 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköping University.