Towards a Corpus of Easy to Read Authority Web Texts

Evelina Rennes, Arne Jönsson

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

Abstract

We present the first version of a corpus of public authorities and municipality web texts, as of spring 2016, divided into easy-to-read texts and texts written in Standard Swedish. The corpus currently contains documents totalling approximately 30 million tokens. In this paper we describe the tools and methods used to collect the web pages and data of the corpus.

1. Introduction

In order to develop techniques for natural language processing, the use of electronically available text resources is crucial. For the task of automatic text simplification, it is valuable to extract patterns and operations from manually simplified texts. There are many techniques and tools available today, but in order to make use of the already existing technologies, more specific digital textual data sources are needed. However, for the task of automatic text simplification, the amount of aligned easy-to-read material in Swedish is limited. Parallel corpora are probably the most useful kind of material for this purpose, but they are often limited, both in number and size. Therefore it is reasonable to also consider comparable corpora (corpora that are similar in content, but lacking perfect alignment) as an alternative beneficial for the different tasks in natural language processing (Santini, 2016).

A corpus of the websites of Swedish authorities and municipalities was created, containing both easy-to-read and standard Swedish material. This corpus is to complement other easy-to-read corpora, most notably the Läsbart corpus (Mühlenbock, 2008), which contains easy-to-read texts, but not their counterparts in Standard Swedish. Other similar initiatives include the Brazilian Portuguese corpus of newspaper articles, where each article was published in two variants (Caseli et al., 2009), and the work on comparable corpora for detecting paraphrases (Deléger et al., 2013). To the best of our knowledge, such aligned material of original and simplified texts does not exist for Swedish, and to harvest the web for easy-to-read material and their parallel counterpart, seems like a reasonable approach to the construction of such a corpus.

2. Method

The collection of the web corpus were roughly conducted according to the following steps:

1. Manual collection of seed URLs
2. Web crawling for collection of URLs
3. Extraction of text from the list of collected URLs
4. Post-processing
5. Preliminary descriptive analysis

2.1 Manual collection of seed URLs

Since the intention was to crawl inside a predefined list of domains, there was no need for constructing query strings for search engines. The crawl was restricted to a number of given domains, and the input consisted of two lists of URLs; one list containing the start pages of the websites, and one list containing the start pages of their easy-to-read counterpart. The collection of the seed URLs was done manually, ensuring that there were no occurrences of erroneous seed URLs.

2.2 Web crawling

The web crawler was built in Python 2.7.0 with Scrapy[1], an open source web scraping framework. The seed URLs consisted of a manually collected list of domains, and the crawler was instructed not to leave the given domain. The municipalities and public authorities that did not have any easy-to-read counterpart were excluded from the crawling.

The crawler started by collecting the easy-to-read versions of each site, in order to be able to differentiate the easy-to-read pages from the rest of the pages. The pages that were not collected during this first round were considered to be in Standard Swedish. The crawler was instructed to respect the robots.txt exclusion directives, and was limited to 1 request per 4 seconds.

2.3 Extraction of text

We used TextCollector (Svensson, 2016) for collecting and filtering the web pages. TextCollector takes a set of URLs as input, collects the pages, and extracts the textual content. A filter excludes broken URLs and any content not written in Swedish.

The corpus material was scrambled at a sentence level, and part-of-speech tagged with Stagger (Östling, 2013), a state-of-the-art tagger for Swedish. The motivation for scrambling the sentences is mainly to avoid copyright issues.

2.4 Post-processing

We are currently working on aligning the material, by matching the easy-to-read text fragments to fragments in Standard Swedish, regardless of where the simple sentence came from. This alignment will initially be carried out

[1] https://scrapy.org/

	Ordinary	Easy-to-read	Total
Number of seed URLs	191	191	382
Number of collected URLs	182,777	2158	184,935
Number of URLs after filterin	181,983	2157	184,140
Raw crawl size	186.1MB	2 MB	188,1 MB
Size after post-processing	2.78 GB	29.1 MB	2.81 GB
Number of pages	136,501	1629	138,130
Number of tokens	29.2M	334,491	29.6M

Table 1: Description of corpus

semi-manually, to create a gold standard, and will then use automatic techniques following an algorithm described in Sanchez-Perez et al. (2014), which original purpose was to detect plagiarism. We have previously used this algorithm on the the SUC and LäsBart corpora (Albertsson et al., 2016). This post-processing will also make the corpus more balanced, as only sentences in the easy-to-read part of the corpus will have corresponding sentences.

3. Descriptive analysis

The total number of URLs collected by the web crawler was 184,935, of which 795 were automatically identified as spam by a filtering script. The 184,140 remaining URLs, of which 2157 consisted of URLs to easy-to-read material, were downloaded by TextCollector. Due to further filtering conducted by TextCollector, the number of documents downloaded were reduced to 138,130, of which 1629 consisted of easy-to-read material.

As can be seen from Table 1 the corpus comprises very few easy-to-read web pages. Furthermore, many pages do not contain useful information; "Under construction" is not uncommon. Further work includes finding such pages, and aligning the easy-to-read pages to their ordinary counterpart. The latter is far from straightforward as most easy-to-read web pages contain information from a number of ordinary pages.

4. Concluding remarks

Textual resources are vital in most natural language processing areas. For the task of automatic text simplification, parallel or comparable material is often used in order to automatically deduce patterns or simplification operations, but the availability of such resources is limited. To harvest the web for easy-to-read texts, and possibly parallel or comparable material, could be a way to effectively construct such resources.

At the time of writing, the post-processing of the collected texts has only begun, and further work is needed to prepare partly parallel, and/or comparable, material. Further improvements will be applied to this first version in terms of a more fine tuned web crawling, text collection and filtering. It is clear that there are many improvement possibilities, such as identifying pages without relevant information, and aligning the easy-to-read pages to their ordinary counterpart. When the corpus is post-processed, and more balanced, we plan to investigate this using readability assessment methods (Falkenjack et al., 2013). When finished, the material will be made freely available.

Acknowledgements

This research was financed by Vinnova and The Internet Foundation in Sweden.

References

Sarah Albertsson, Evelina Rennes, and Arne Jönsson. 2016. Similarity-based alignment of monolingual corpora for text simplification purposes. In *Proceedings of the Coling Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), Osaka, Japan*.

Helena. M. Caseli, Tiago. F. Pereira, Lucia. Specia, Thiago. A. Pardo, Caroline. Gasperin, and Sandra. M. Aluisio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. In *Proceedings of CICLing*.

Louise Deléger, Bruno Cartoni, and Pierre Zweigenbaum. 2013. Paraphrase detection in monolingual specialized/lay corpora. *Building and Using Comparable Corpora*.

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.

Katarina Mühlenbock. 2008. Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In Anju Saxena and Åke Viberg, editors, *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for swedish. *Northen Europea Journal of Language Technology*, 3.

Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh. 2014. The winning approach to text alignment for text reuse detection at PAN 2014: Notebook for PAN at CLEF 2014. *CEUR Workshop Proceedings*, 1180:1004–1011.

Marina Santini. 2016. Bootstrapping domain-specific comparable corpora from the web: The case of swedish "myndigheter". Technical report, SICS East Swedish ICT AB.

Cassandra Svensson. 2016. Creation and evaluation of TextCollector. Technical report, Linköping University.