Semiautomatic Simplification to Improve Readability of Texts for People with Special Needs

Susana Bautista

Departamento Ingeniería del Software e Inteligencia Artificial,
Facultad de Informática,
Universidad Complutense de Madrid, Spain
subautis@fdi.ucm.es
Supervisor: Dr. D. Pablo Gervás

Resumen The way of writing or presenting the information can exclude many people, especially those who have problems to read, write or understand. The process of simplifying texts by hand is extremely time and effort consuming. This work presents a semiautomatic textual simplification system to help in this adaptation process from original texts to their easy-to-read versions. Working with a parallel corpus generated by experts, we use a subset of the corpus to train and identify different transformations in order to automate them. The rest of the corpus is used as test corpus to check the success of the automatic transformations.

Key words: simplification, adaptation, accessibility, readability

1. Motivation

The Standard Rules on the Equalization of Opportunities for Persons with Disabilities of United Nations require that all information services and all documents are adapted so they will be easily accessible to everybody [1]. The aim of this work is the development of an architecture for the semiautomatic and interactive generation of easy-to-read texts. Our system aspires to assist people who have to create easy-to-read versions of texts by automating some transformations (for example, syntactic and lexical) which appear always in this kind of texts.

2. Work Carried Out so Far

Previous work has been oriented towards exploring syntactic simplification for easy-to-read texts [2]. These efforts were driven by existing European guidelines on how to achieve an easy-to-read text.[1] These guidelines were use as starting point to define syntactic transformation rules over syntax trees obtained

[1] http://www.osmhi.org/contentpics/139/European %20Guidelines %20for %20ETR %20publications %20 %282 %29.pdf

using the Stanford Parser [3]. The transformation rules produce a simplified syntax tree from which a simplified version of the original sentence can be reconstructed. Evaluation of the performance of the system was carried out using existing readability metrics [4]. These metrics are based on correlative measures of particular elements of the text, and output a numerical value indicating the number of years of formal education required to easily understand the text after the first reading. Relying on these metrics to measure the improvement of readability texts, we observe that the level of difficulty of the text decreases. The improvement is significant even with the small set of rules that was used.

This approach has several shortcomings. First, readability metrics cannot measure how comprehensible a text is, since text comprehension not only depends on text features, but also on readers characteristics (prior knowledge, reading and comprehension abilities, and so on). And they cannot measure whether a text is suitable for particular readers needs [5]. Second, there is no guarantee that no relevant information is lost during the applied transformations. Third, the guidelines used as inspiration, though generally useful, proved to be insufficient to establish clear criteria for identifying the full range of possible successful substitutions. Fourth, syntactic transformation only covers a small subset of the possible operations that human editors carry out during adaptation of texts for easy reading.

In view of these results, it was deemed necessary to find an alternative source to identify the set of possible operations and their relative frequency, and from which to obtain appropriate criteria to drive the process. To this end, a corpus-based approach is applied. Corpus-based approaches have become very popular in natural language processing over the last decades. In the particular case of Machine Translation (MT), pairs of translated sentences from a bilingual corpus are aligned, and occurrence patterns of words in two languages in the text are extracted and matched using correlation measures. Besides, MT systems are generally trained using sentence-aligned parallel corpora. This methodology developed in MT is used for different kinds of applications: extracting paraphrases from a parallel corpus [6] or incorporating word-level alignments into the parameter estimation of models in order to reduce alignment error rate [7].

The case of transformation of texts into easy-to-read versions can be phrased as a translation problem between two different subsets of language (the original one and the easy-to-read version). This paper proposes the application of a corpus-based solution to the problem phrased in this way.

3. Approach for Semiautomatic Simplification of Texts

The system proposed is intended to help in the adaptation of texts to easy-to-read versions to increase the availability and coverage of this type of material.

As source corpus we have chosen a subset of the documents available in the web page of Inclusion Europe[2]. In this web there are two versions of each page:

[2] www.inclusion-europe.org

the original and the easy-to-read. With these pages we can build a parallel corpus of texts, part of which will then be used as training corpus for the transformation process and part as test corpus. Although the web page is translated to all the official languages of the European Union, we will start working with the English version.

3.1. Corpus Elaboration

In order to be useful, the corpus of parallel texts has to be processed and annotated. In contrast to applications for translation, the corpus we are interested in will involve operations such as deletion, rewriting, and possibly addition of explanatory information. The construction of the corpus involves identifying the correspondences between the sentences in the original text and the sentences of the simplified text. This corresponds to establishing alignment at the level of sentences. There is a second level of processing that corresponds to annotating each sentence with its syntactic structure. At this level, it is also necessary, for sentences which correspond to one another, to identify correspondences between parts of the original sentence and parts of the simplified version. This will occur both at the level of syntactic structures (correspondence between syntactic subtrees) and lexical items (correspondence between specific words that have been replaced with others).

The corpus so processed becomes a very powerful source of knowledge. First, it provides very important information regarding the type of operations carried out during adaptation, and the relative frequencies of each type of operation. Second, it constitutes a valuable resource for identifying empirical instances of the various types of operation, which can be used as data during development or even training of an automatic or semi-automatic system.

3.2. Automatization of the Process

Once the annotated corpus is available, the general schema for the subsequent process is as follows:

1. A subset of the operation types identified above is selected, based on their relative frequency of use and the availability of computational resources for automating them.
2. For the particular operations selected, an empirical process of rule extrapolation is carried out over the set of correspondences for the relevant operation contained in the subset of the corpus set aside for training.
3. A computational implementation of these rules is developed.
4. The resulting module is applied to the source texts corresponding to the part of the corpus set aside for testing, and results are compared to the associated output versions.

Based on the work carried out so far along these lines [2], we reckon that the most likely operations to consider as candidates for successful automation are syntactic transformation of parse trees and lexical substitution.

For syntactic transformation of parse trees we are considering the use of the Stanford Parser [3] to obtain the syntactic structure of the input text. This will ensure that the syntax trees obtained will match in format and notation those in the corpus, which will ensure applicability of the transformation rules.

For lexical substitution, we are considering the use of WordNet [8] as resource. Preliminary work on this approach has shown that WordNet provides possible candidates for substitution based on different criteria (more frequently used synonyms for a given word, less ambiguous synonyms for a given word, more specific hyponyms when the original word is too abstract, or more abstract hypernyms when the original word is too specific). Although some of these possibilities show advantages (simplified texts with lower scores in terms of readability metrics, or with a reduced count of ambiguous words), it was unclear at the end of the day which criteria should be preferred to obtain more useful simplified texts. It is expected that the application of the corpus approach will provide clear guidelines as to what type of substitution to apply in different situations.

Referencias

1. Nations, U.: Standard Rules on the Equalization of Opportunities for Persons with Disabilities. Technical report (1994)
2. Bautista, S., Gervás, P., Madrid, R.: Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In: The Second International Conference on Information and Communication Technologies and Accessibility. (2009)
3. Klein, D., Manning, C.D.: Fast Exact Inference with a Factored Model for Natural Language Parsing. In: In Advances in Neural Information Processing Systems 15. (2003) 3–10
4. Kincaid, J.P., Fishburne, R. P., J., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Behaviour Research Methods (1975)
5. Zakaluk, B., Samuels, S.: Readability: Its past, present, and future. International Reading Association (1988)
6. Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2001) 50–57
7. Callison-Burch, C., Talbot, D., Osborne, M.: Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora. In: ACL. (2004) 175–182
8. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. Int J Lexicography **3**(4) (1990) 235–244