Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability

Susana Bautista
Dep. Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid
Madrid (Spain)
Email:
subautis@fdi.ucm.es

Pablo Gervás
Instituto de Tecnología del Conocimiento
Universidad Complutense de Madrid
Madrid (Spain)
Email:
pgervas@sip.ucm.es

R.Ignacio Madrid
R & D Department
TECHNOSITE
Fundosa Group
ONCE Foundation
Madrid (Spain)
Email:
nmadrid@technosite.es

Abstract[1]

WCAG 2.0 (W3C Recommendation 11 Dec. 2008) recommendation includes a specific guideline regarding the readability of text contents. This paper presents an approach to adapt text to improve readability. We use existing guidelines for the development of easy to read materials together with natural language processing (NLP) methods to adapt text for people with learning or comprehension difficulties. The simplifications introduced address two basic problems: complex syntactic structure, and use of difficult lexical terms. Syntax structure is handled by obtaining syntactic parse trees of the sentences and applying a set of conversion rules to simplify them. Lexical complexity is addressed by replacing difficult words with simpler synonyms obtained from an electronic lexical database. We rely on readability metrics provided by existing online programs to identify problems and validate tentative solutions. After processing, the texts have been converted to accommodate the characteristics of people with special comprehension needs.

1. Introduction

The ability to access information about their society's culture, literature, laws, local and national policies and ethos is fundamental for people to be able to take part in mainstream life. However, most textual information currently available is presented in such a way as to be very difficult to understand for people with limited skills in reading, writing or understanding. People can have problems with understanding written text due to learning or cognitive disabilities, lack of sufficient formal education, social problems or simply because they are immigrants whose mother tongue is a different language. The need for simple procedures to obtain easy to read versions of existing material has become paramount with the recently approved Web Content Accessibility Guidelines 2.0[1]. The

[1] The research described in this paper/work is partly supported by the Spanish CDTI's project INREDIS (CEN-2007-2011) [www.inredis.es], under the INGENIO 2010 programme. The opinions expressed in this paper are those of the authors and are not necessarily those of the INREDIS project's partners or of the CDTI.

new version of W3C standard for accessible Web contents includes as one of its explicit guidelines (3.1.) that text content must be readable and understandable.

Natural language processing technology may provide the means for semi-automatically obtaining easy-to-read versions of documents, thereby improving their accessibility and reducing the workload of converting them.

An easy-to-read document can be defined as one that contains only the most important information written and presented in the most direct way so that the largest possible audience can understand it. The way in which a document is structured is very important. The contents should follow a clear and logical sequence. All unnecessary ideas, words, sentences or phrases should be avoided or removed. The presentation of the information is also very important. Photographs, pictures or symbols should support the text wherever possible in order to aid understanding.

Our approach is based on state-of-the-art readability models [21] and follows guidelines on how to produce texts and summaries which are easy to read and understand. The objective is to test the feasibility of a semi-automatic method to improve the readability of texts.

2. Previous work

In this section we present the tools we apply in our approach: the guidelines about easy to read materials, the readability metrics, the lexical database, and the natural language parsing tool.

2.1. Guidelines

The ILSMH European Association, [9] has undertaken a project to develop "Guidelines for the production of easy to read materials". Easy-to-read materials are also easy to comprehend. Because they get the message across clearly, they benefit everyone, not just people with literacy problems. The concept of "easy-to-read" can not be universal and it will not be possible to write a text that will suit the abilities of all people with literacy and comprehension problems. The set of guidelines presented by ILSMH is summarised in Table 1. When we write documents adapted to the needs of the final user, we have to bear in mind the following rules to avoid fixed constructions and fixed words to improve the readability of the text.

Use simple, straightforward language	Avoid abstract concepts
Use short everyday words	Use many personal words
Use practical examples	Address the readers with respect
Use short sentences mostly	Only one main idea per sentence
Use positive language	Use active rather than passive verbs
Do not assume previous knowledge	Use words consistently
Keep the punctuation simple	Do not use the subjunctive tense
Avoid unusual metaphors	Be careful with numbers
Do not use foreign words	Avoid cross references
Mention a contact address	Avoid jargon and abbreviations

Table 1: Summary of Easy to Read Guidelines

2.2. Readability Metrics

Readable is defined as “fit to be read, interesting, agreeable and attractive in style and enjoyable”. From the earliest efforts to the present day, readability tests have been designed as mathematical equations which correlate measurable elements of writing - such as the number of personal pronouns in the text, the average number of syllables in words or number of words in sentences in the text. There are different metrics designed to measure the readability of a sample of English writing. The resulting number is an indication of the number of years of formal education that a person requires in order to easily understand the text on the first reading. These metrics are sometimes referred to as tests and sometimes as formulas. We refer to them using the encompassing term of metrics because it reflects better the role they play in our approach.

Obviously, readability metrics cannot measure features like interest and enjoyment. Also, when we ask whether text is understood by its reader we are questioning its "comprehensibility". 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 metacomprehension abilities, and so on). And they cannot measure whether a text is suitable for particular readers needs. [21]

There are two metrics, the **Flesch Reading Ease** [13], and the **Flesch–Kincaid Grade Level** [11] which use the same core measures (word length and sentence length). Both metrics were devised by Rudolf Flesch. In the Flesch Reading Ease metric, higher scores correspond to materials that are easier to read.

Unlike the other metrics, the **Automated Readability Index** [2], along with the **Coleman-Liau** [5], relies on a factor of characters per word, instead of the usual syllables per word. Although opinion varies on its accuracy as compared to the syllables/word and complex words indices, characters/word is often easier to calculate, as the number of characters is more readily and accurately counted by computer programs than syllables.

Finally, the **Gunning Fog Index** [14] is computed as the average number of words per sentence. **SMOG** (Simple Measure of Gobbledygook) [18] was published by G. Harry McLaughlin in 1969 as a more accurate and more easily calculated substitute for the Gunning-Fog Index.

2.3. Electronic Lexical Resources

Lexical resources are a set of electronic corpora, lexicons and dictionaries for use in natural language processing. Over the last decade their use as support tools for Natural Language Processing and Information Retrieval has expanded significantly, and the number of resources available has grown. Some important examples are: Roget’s Thesaurus [17], FrameNet [12], Extended WordNet [10], VerbNet[19], Lexical Conceptual Structures [15] and WordNet[7]. For the purpose of our work, we required a resource capable of providing synonymy information over a large number of words. For this reason, we have focused on WordNet [7], a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Different senses of a word are in different synsets. The

synsets are organised into a taxonomy by means of hypernym/hyponym links that identify abstractions and specific instantiations of concepts respectively.

2.4. Natural Language Parsing

A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s. Classic parsers such as the Charniak Parser [4], Collins Parser [6], or Bikel Parser [3] are now recognised as milestones that have made it possible to achieve the current standards of accuracy and efficiency. The more recently developed family of dependency parsers (such as Minipar [16]) provides similar functionality by abstracting away from the relative position of words in a sentence and focusing on the dependencies between them. For the purposes of the work reported here, traditional constituent grammars are better suited (because we do need to consider the relative positions of words when transforming sentences). Therefore we have chosen the Stanford Parser [8], a freely available statistical parsing system. It provides good performance and it is reasonably efficient.

3. Steps for a Semiautomatic Conversion

The ILSMH guidelines summarised in Table 1 are intended for authors about to produce text, not for editors intending to convert original text into and easier to read version. For this reason, they are phrased in a very generic way. Some of them concern the kind of content to be included (“Use practical examples”). It is clearly beyond the scope of our approach to add practical examples when they are not present. In general, we focus on a subset of these guidelines that lend themselves to be supported by the tools we are considering. Among these, we consider guidelines that can be addressed by means of syntactic transformation (“Use simple, straightforward language”, “Use short sentences mostly”, “Only one main idea per sentence”) and some that can be addressed by means of lexical substitution (“Use short everyday words”, “Avoid abstract concepts”).

3.1. Syntactic Conversion

With respect to the first subset of the guidelines, we apply the following procedure. A readability measure is obtained for a given sentence by submitting it to an interactive web page [20]. A syntax tree for the given sentence is obtained from the Stanford parser [8]. The resulting syntactic structure is checked for complex subtrees that might be simplified. Figure 1 shows the original tree for sentence: *“He’s from Madrid in Spain but he lives in Buenos Aires in Argentina”* and the simplified version of the tree after two applications of the conversion rule to transform the subtrees that describe place of origin and place of residence. The resulting sentence is: *“He’s from Spain but he lives in Argentina”*.

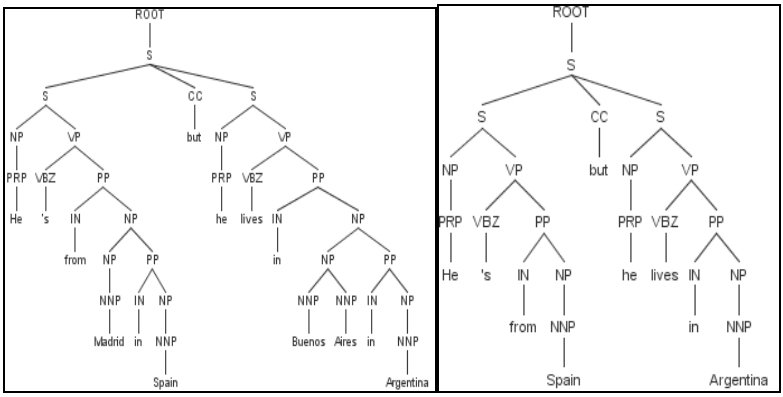

Figure 1: Output of the Parser

If a simplification is identified, it is described in terms of a conversion rule from the complex version of the subtree to the simplified version of the subtree. The version of the sentence that results from applying this simplification is resubmitted to the interactive web page to obtain a readability measure. If this measure is an improvement on the earlier values, the conversion rule for this syntactic simplification is retained for future use by the system. Table 2 shows examples of conversion rules obtained in this manner.[2]

1. NP(DT,NN) → NP(DT)
2. PP(IN, NP(NP(NNP),PP(IN, NP(NNP)))) → PP(IN, NP(NNP))
3. S(NP,VP(VBZ,NP,SBAR(IN,S(NP,VP)))) → S(NP,VP)
4. VP(VBZ,NP(NP(CD),PP(IN,NP)))→ VP(VBZ,NP(CD))
5. VP(VBZ,NP,PP)→ VP(VBZ,NP)

Table 2: Some rules of conversion of a text

These rules have no direct correlation with the guidelines in Table 1. Although each conversion rule could probably be classified as an instance of the observance of one of the guidelines, conversion rules will be specific to a particular syntactic construction (so they might be paraphrased verbally as "eliminate superfluous adjectives", or "split conjunctions of sentences into two different sentences").

3.2. Lexical Substitution

With respect to the subset of the guidelines susceptible of being addressed by means of lexical substitution, we apply the following procedure. Having obtained an initial readability measure, we identify those words that are likely to affect the readability scores more highly. Given the metrics we are employing, these will be either words with many syllables or with many characters. We look up these words in WordNet [7] to obtain a set of synonyms. Among these, we select a single term by choosing the one that is shorter in

[2] More complex rules (complex sentence→simple sentence1 + simple sentence2) may be considered as future work

terms of characters or syllables. We replace the original word with the selected term in the original sentence. We iterate over the set of words in the sentence.[3]

WordNet is structured as a taxonomy of concepts. Each concept is linked to its hypernym (the next concept up in the taxonomy). When we carry out the replacements using WordNet, we can also consider the hypernyms of the word. So, in the example of the text, we can replace "enormous" by "big" because it is a hypernym of the word "enormous". In such cases, the results do need to be validated by resubmitting the sentence to the interactive web page. Only hypernym substitutions so validated are retained.

3.3. An Example of Text Conversion

Figure 2 shows an example input original text together with final text and Table 3 shows the results for the selected metrics after each process of conversion.

This person is Juan Uribe. He's from Madrid in Spain but he lives in Buenos Aires in Argentina. He is a spectacular actor. Mr Uribe is a young person, he is 23 years old. He works for a famous TV Argentina. He also studies Journalism at the Federal University of Argentina. Mr Uribe cannot sing o dance well but he can play the violin very well. He loves baroque classical music and he listens to it every day from 5:10 am to 8.45 am every morning. After this he has a big glass of fruit juice for breakfast and goes to walk to university. Mr Uribe is a very good student. He is an intelligent person because he speaks several languages. He speaks Spanish, Portuguese and English. Mr Uribe lives in an enormous flat. His flat number is 133, on the 1 st floor. Mr Uribe loves Spanish food as paella but he doesn't like Argentine food. Something so typical as dulce de leche. He loves milk caramels but he doesn't like chocolate. He doesn't smoke. He hates it.

This is Juan Uribe. He's from Spain but he lives in Argentina. He is an actor. Mr Uribe is 23 years old. He works for TV ARGENTINA. He also studies; he studies Journalism at the Federal University of Argentina. Mr Uribe cannot sing or dance but he can play the violin very well. He loves classical music and he listens to it every day from 5:10 am to 8:45am every morning. After this he has fruit juice for breakfast and goes to university. Mr Uribe is a very good student. He speaks Spanish, Portuguese, and English. Mr Uribe lives in a big flat. His flat number is 133. Mr Uribe loves Spanish food but he doesn't like Argentine food. He loves milk caramels but he doesn't like chocolate. He doesn't smoke.

Figure 2: Example text

Text	FRES	ARI	FKGL	CLI	GFI	SMOG
T	77.2	3	4.8	5.7	8.2	8.6
T+WN	78.2	2.9	4.7	5.6	8	8.5
T+P	78.2	2.4	4.3	5.4	7.7	8.1
T+WN+P	81.1	2.3	3.9	5.1	7.3	8

Table 3: Outputs of the readability tests

(T: original text; +WN: lexical substitution applied; +P: syntactic conversion rules applied)

Although only simple examples have been considered, we can see how with a single lexical substitution using WordNet (i.e. we replaced one adjective by another shorter synonymous, in this case the word *enormous* was replaced by *big*) the results improved. When we use some rules of conversion to achieve easy to read texts, the outputs of the tests show an improvement in the readability of the text according to more than one metric.

[3] Since the process by definition reduces the complexity of the words, we have considered it unnecessary to validate these changes by obtaining a second readability measure.

The results need to be compared with adaptation by human editors to identify how meaningful the improvements are. Until these improvements have been validated the process should not be fully automated.

4. Discussion

The procedure described above is validated by the readability metrics. However, this is no guarantee that all the guidelines have been observed to the same extent. For instance, the substitution of a long word by a shorter hypernym results in an improvement of the readability metrics (concerned as they are with only the length of words), but it constitutes a transgression of the guideline "avoid abstract concepts" (the hypernym of a word is by definition more abstract than the word itself).Another example of conflicts between different guidelines arises from the processes of syntactic transformation. Some of the simplifications applied to reduce the syntactic complexity of a sentence may in fact eliminate practical examples that were included in the original text, thereby constituting a transgression of the guideline "use practical examples". If we observe Table 3, we could see how the results of the FRES test [13] increase with the changes, improving the readability of the text. The rest of outputs of the test refer to the school grade required to understand the text. So, we can see that the outputs of the ARI [2] and CLI [5] tests are lower than the rest. These tests count the number of characters per word instead of counting the number of words per sentence as the rest of the tests. Because of this, the outputs of the tests GFI [14] and SMOG [18] are higher. If we look at the output of the SMOG test, it indicates that formal education up to the 9^{th} grade is required to understand the original text and it has decreased to 8^{th} grade for the final text. It suggests that the changes are indeed achieving some reduction of the reading difficulty of the text. A final point to consider is that the resulting texts would still have to be tested to ensure that no critical information is being lost during the conversion process. This issue has not been considered in the present paper, but it would provide fundamental information for a decision on the practical applicability of these processes. To obtain this information, experiments with real users, possibly based on practical tasks, would have to be conducted to measure the amount of information loss.

5. Conclusions and Future Work

The present paper outlines preliminary work aimed at identifying whether automatic processes of conversion based on natural language processing techniques might succeed in providing easy to read fast drafts of given texts. The proposed processes of conversion have been applied to a small set of example texts, and validated numerically by the application of a suite of readability metrics. The results show marked improvements in terms of the readability metrics. Nevertheless, two issues have been identified that would have to be further studied before a complete conclusion can be reached. First, only a subset of the guidelines is actually addressed by the readability metrics, those relating to syntactic or lexical characteristic of the text. Other metrics are concerned with the semantic content of the text, which is not addressed by these metrics. This suggests two possible extensions: the proposed processes should be refined to ensure they respect the guidelines concerned with semantic content, and additional processes may be included which address those semantic guidelines specifically. The set of guidelines used as inspiration has been shown to include conflicting guidelines. Second, the simplification process proposed does not

consider the possible information loss incurred in each step. This may results in easy to read versions of the text that do not actually convey the intended information. Additional testing procedures must be included to ensure that information loss is measured, and taken into account in the validation of the proposed transformation processes. These testing procedure are likely to include experimental validation by users.

Once the full set of processes – including both lexical-syntactic and semantic information – and the full set of validation mechanisms – including readability metrics and information preservation – have been identified they should be tested systematically over a large set of texts. The results of these tests should be employed to refine the definition of the processes. As a future goal, we would like to automate the process of achieving text generation for users with different reading needs. Extensions to others languages may be considered subject to availability of parsers and lexical resources.

7. References

[1]. Web Content Accessibility Guidelines 2.0 WC3 Recommendation 11 December 2008, http://www.w3.org/TR/WCAG20/ Last Visited: February 2009

[2] ARI: http://www.oleandersolutions.com/ari.html Last Visited: November 2008

[3] D. M. Bikel. 2004. Intricacies of Collins' Parsing Model. In Computational Linguistics, 30(4), pp. 479-511.

[4] *Charniak, E. Statistical parsing with a context-free grammar and word statistics*, Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI Press/MIT Press, Menlo Park (1997).

[5] CLI: http://www.oleandersolutions.com/colemanliau.html Last Visited: November 2008

[6] M. Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. Proceedings of the 34th Annual Meeting of the ACL, Santa Cruz.

[7] C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database, MIT Press, 1998. Website: http://wordnet.princeton.edu/

[8] D. Klein and C. D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.Website: http://nlp.stanford.edu/software/lex-parser.shtml

[9] European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability, http://www.inclusion-europe.org/uploads/doc/99.pdf

[10] Extended WordNet: http://xwn.hlt.utdallas.edu/ Last Visited: November 2008

[11] FKGL: http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php

[12] FrameNet: Atkins, S., M. Rundell and Hiroaki Sato (2003). The Contribution of Framenet to Practical Lexicography, International Journal of Lexicography, Volume 16.3: 333-357.

[13] FRES: http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php

[14] Gunning-Fog Index: http://www.readabilityformulas.com/gunning-fog-readability-formula.php

[15] LCS: http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

[16] Lin, D. 1998. Dependency-based evaluation of MINIPAR. In Proc. of Workshop on the Evaluation of Parsing Systems, Granada, Spain.

[17] Roget's Thesaurus: http://www.nzdl.org/ELKB/ Last Visited: November 2008

[18] SMOG: http://www.readabilityformulas.com/smog-readability-formula.php

[19] VerbNet: K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extending VerbNet with Novel Verb Classes. Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. June, 2006.

[20] Website: http://www.editcentral.com/gwt/com.editcentral.EC/EC.html

[21] Zakaluk, B.L. and S.J. Samuels, Readability: Its past, present, and future. 1988, Newark: International Reading Association.