SOFTWARE FEASIBILITY STUDY TO TRANSFORM COMPLEX SCIENTIFIC WRITTEN KNOWLEDGE TO A CLEAR, RATIONALE AND SIMPLE LANGUAGE

Manoj Khandelwal[1] and Mohammad Jafarabad[2]

[1]Faculty of Science and Technology, Federation University Australia, PO Box 663, Ballarat, Victoria 3353, Australia
Phone: +61 3 5327 9821
[2]Department of Computer Engineering, Young Researchers Club, Robatkarim Branch, Islamic Azad University, RobatKarim, Iran.

ABSTRACT

Ali Akbar Dehkhoda, the prominent lexicographer, describes a person who has difficulty in grasping knowledge as someone who "Cannot understand something without knowing all its details." If the knowledge required by somebody is in a language other than the person's mother tongue, access to this knowledge will surely meet special difficulties resulting from the person's lack of mastery over the second language. Any project that can monitor knowledge sources written in English and change them into the user's language by employing a simple understandable model is capable of being a knowledge-based project with a world view regarding text simplification. This article creates a knowledge system, investigates some algorithms for analyzing contents of complex texts, and presents solutions for changing such texts simple and understandable ones. Texts are automatically analyzed and their ambiguous points are identified by software, but it is the author or the human agent who makes decisions concerning omission of the ambiguities or correction of the texts.

KEYWORDS

Written knowledge, complex texts, text simplification, ambiguous words

1. INTRODUCTION

Knowledge that is implemented and published in written form is created correspondent to the literacy level of its author. Previous studies, the extent of proficiency in writing techniques, and the viewpoints of the author of any knowledge, play an important role in the process of designing and implementing documents related to the knowledge. In most cases, before written knowledge (created by an expert in the related industry or a scientist) is published, it passes through a process of correcting the writing of or the scientific errors in the text by the editorial board of the journal, conference, or publications. Moreover, the educational level or the areas of study the readers of the texts must have sometimes specified in the classification of some forms of

knowledge. The problems that will then arise for the readers are ambiguities and complexities in the texts that make them difficult to understand.

Understanding knowledge written by a specialist is not an easy task for beginners because they will need great mental activity and intelligence to adapt their level of knowledge to that of the specialists. Difficult expressions or complex sentences containing unfamiliar words can prevent grasp of the written knowledge or lead to misunderstanding it. Special arrangements are required to prevent this. Moreover, if knowledge is to be written in a way as to follow all the details, conventions, methods, and rules of formal writing, it will entail great costs for the organization publishing this kind of knowledge. Therefore, projects for disambiguation of knowledge containing texts must be executed in a way that most of the ambiguities present in these texts are omitted at minimal cost.

Authors, scholars, and researchers present their personal knowledge through theoretical-applied research articles in publications related to their field of study, or in book form, to be used by the public. Disambiguation of educational texts can be performed through various processes. Knowledge managers and users, or intelligent software, can discover ambiguities in texts. The creator of the knowledge can also carry out correction of text ambiguities and complexities at minimal cost. In this article, the iterative and augmentative method of software production is studied and analyzed to execute the project for identifying ambiguities and literary complexities in knowledge containing texts. The project will make it possible to implement software that can identity ambiguous words or sentences and, in addition, play a considerable role in removing these ambiguities through making suitable recommendations to the author of the knowledge containing text.

2. BACKGROUND OF TEXT SIMPLIFICATION

Language is one of the most important of mental phenomena and has been in use for many years among people living in various regions. At present, there are about 7000 languages in the world, 2400 of these are on the verge of extinction, and 840 are used only in New Guinea. Linguists investigate languages in the domains of syntax, phonology, morphology, and semantics. Syntax deals with sentence recognition and studies the rules by which words are combined and placed next to each other. Phonology is concerned with the positions of chain and sub-chain phonetic elements in the language system. Morphology is a part of language grammar that analyzes the structure of words by recognizing their morphemes), while semantics studies meanings in human languages.

The famous linguist Darrow has said, "There is a completely explicit, clear, and standard-based relationship with great importance between language and its content, and knowledge, cognition, and language are the most important constituents of learning" [9]. The necessary condition for reducing the time it takes to understand the content of a written text, and to increase the speed of learning, is for the text to be fluent, simple, and without ambiguities. If a text is written fluently, understanding the text and correct thinking about the subject happen simultaneously with reading it. If a text is complicated and difficult to grasp, only partial thinking about the subject takes place during the reading, and the reader completes the thinking process by asking questions. The effort required for finding answers to these questions makes the reader less inclined to receive the knowledge (and this is inconsistent with the principles of knowledge management)

Due to the complexity of language structure, most text processing systems do not achieve satisfactory success when put to use. Google Website Translation is among the most successful text processing systems. Another site active in content analysis is the Online Summarization Site. These sites apply their proprietary algorithms to analyze texts. So far, no successful online or offline software has been successfully applied to turn complex texts into simple ones. However, a series of research has been conducted by a team of researchers at Lexile Company that is presented in the form of Lexile Framework.

Lexile offers a standard called LF that evaluates the difficulty of reading texts and the capacity of individuals in reading them. The Lexile Level ranges from 0 to 1800, and each book is assigned one of the numbers in this range. The difficulty of reading texts is measured based on a series of mathematical formulae. Results of these calculations are obtained based on two features: semantic difficulty based on words of the same family with respect to frequency of use in the text, and syntactic complexity based on sentence length [2].

After the evaluations, the LF assigns a value to the difficulty of the text (Table 1). Based on the Lexile indices, and considering the age and education level of each individual, it will be possible to determine the capacity of each person in reading the text.

Table 1: Lexile standard

Rank	Standard of the difficulty of texts	Standard of the reading capacity of individuals
1	200-400	Up to 300
2	300-500	140-500
3	500-700	330-700
4	650- 850	445-810
5	750-950	565-910
6	850-1050	665-1000

The LF recognizes the complexity of texts but has two main shortcomings. First, the whole text is not processed, only one or a few sample paragraphs from the text are analyzed and, based on the analysis, the difficulty of the whole text is shown by a score. Second, LF does not specify the reasons for the difficulty of the text, while in the idea proposed by us the reasons for the difficulty of the text must be specified for its author. Other research on text simplification has been carried out as case studies and in the three areas of mathematics, computers, and literature. Among the most important of them are various projects for identifying the role played by pronoun antecedents in sentence complexity and the study of ambiguity in words that are spelled the same but have different meanings.

3. STUDY OF THE IMPORTANCE AND THE GOAL OF SIMPLIFYING TEXTS WRITTEN IN ENGLISH

English plays a basic and important role in the daily lives of people because it has become a communication bridge between various societies and most international communications are carried out in this language. Moreover, most reference books in various sciences are written in English and specialists and students studying at various levels and in different fields need to have sufficient knowledge of English to be able to use these references. Therefore, unfamiliarity with this language causes problems in carrying out various duties, one of the most important of which is the impossibility of receiving up-to-date and needed knowledge.

The main purpose in presenting this project of simplifying texts was to make reading texts in English possible for those having university education but low levels of acquaintance with the English language, although execution of this project can be useful, in many instances, for other sectors of the society. Since one of the main mottos of the United Nations Organization UNESCO is eradication of illiteracy in societies, simplification of knowledge containing texts can also be considered one of the tools of realizing the high goals of this Organization. In this section, two main classes of people are studied as target groups for easy reading of science containing texts.

3.1. Those with low capacity in language learning

Research has shown more than 25% of the population in the Unites States does not acquire the expected level of reading ability after 9 years of formal education. This figure reaches 45-50% in several other countries. Illiteracy and inability to apply the basic vocabulary and comprehension skills learned in reading techniques may be due to various factors such as insufficient education, non-standard education, social problems, mild mental retardation, etc. People with such problems are not able to read and understand applied texts available in advanced societies.

The goal in this article is to prepare the ground for faster comprehension of scientific texts by people with university education. As expected, a native speaker of English with this level of education is capable of sufficiently understanding such texts, but will certainly face difficulties when reading a text in which the standard skills of the English language have not been observed. The reason for this is that having academic education and being native speakers are not sufficient conditions for understanding every book written in a special field. Experience and continuous studies are other required conditions for comprehending texts, and they will help readers to read complex texts.

3.2. Non-native speakers

Any language learned after the first one (the mother tongue) is called "second language." Eric Linenberg, a pioneering linguist and neurologist, called second language a language consciously learned or spoken by a person after maturity. In most cases, people never reach a similar level of fluency and comprehension in the second language as their mother tongues, and face problems in finding correct words and grammatical structures. Learning a second language can be a life-long process for many people. Psychologists have reached the conclusion the reason for this is that there are separate centers in the brain for learning and for controlling skills in a second language. It must also be noted that people weak in learning a new language are different from the mentally retarded.

There are 1.3 billion people living in China and, in many cases, they access the required knowledge only in sources that are in Chinese. Therefore, there is no great need to learn English in China. Since English has been omitted from college entrance examinations, studying a scientific text in English for graduates of the best universities in China is a difficult and costly task [8]. Our proposed project to simplify scientific texts written in English, and to make them understandable for people with low levels of proficiency in this language, can substantially help this section of the world population.

Understanding written text for the blind is like learning a foreign language. These people have limited understanding of textual metaphors. However, blind people can use easy-to-read materials in which the Braille form and format is employed. A blind French person learns French Braille and, hence, using easy-to-read materials helps this individual to read and understand texts written in English Braille [1].

4. LEVELS OF SIMPLIFYING TEXTS

The need students and researchers have for reading specialized texts, and the various levels of familiarity with the language, make it more difficult to set a defined level for simplifying texts. There are various levels, classes, classifications, and models in each language. Considering the level of language skills a person has depends on many factors such as age, occupation, level of education, personal interests in the language, talent, etc., classification of people and mentioning evidence for each various language level is difficult and, sometimes, seems to be impossible. Therefore, examples are often not presented for each level. For example, it is not possible to say what level of language proficiency a certain age group or all graduates of a given level of education have reached. Therefore, complex study and research are needed to assign a level of language proficiency to a group of people with one or more common characteristics [9].

Determining the level of English language skills in people is a very complex process. Various models have been designed for classifying the levels of the English language. These models are conventional and divided into the three classes of advanced, intermediate, and Beginner levels. The advanced level is a high level of language proficiency in people possessing the capabilities of the English language. People whose mother tongues are not English but have attained an advanced level of proficiency in English and speak very much like native speakers are called near native speakers. These people cannot be easily differentiated from native speakers.

People with the advanced level of skills in reading English texts can study and understand English texts and do not need text simplification, although text simplification accelerates their comprehension. In some cases, the presence of ambiguous and complex sentences in texts makes their understanding difficult, even for native speakers. In these cases, only a linguistics specialist can analyze these texts. Therefore, although text simplification will greatly help people with intermediate or beginner level proficiency in the language, in many cases it will certainly solve problems native speakers or people with advanced levels of proficiency in English face when reading ambiguous or complex texts.

5. DESIGNING THE PROPOSED SOFTWARE SYSTEM

Considering various studies have been carried out for many years on correcting the grammars of various languages, and especially that of English, with the goal of removing literary ambiguities, all grammatical points of the language have been presented in thousands of books. Therefore, the

purpose of this project is not to carry out new linguistic studies, but rather previous studies are analyzed to present an intelligent system for discovering the main factors contributing to text complexity through changing the existing language rules into algorithms. Obviously, accurate recognition of all factors influencing text complexity requires complex text processing systems and, if facts regarding software and hardware infrastructure are not considered in the feasibility study, problems will arise when implementing this project. Therefore, the whole idea behind this project will be to use new algorithms within the scope of previous linguistic studies.

The proposed project includes producing software for analyzing texts and for finding ambiguities in them. The software must read the text and then point out the ambiguities and complexities to its author. The author can then edit the text and omit the challenging points by paying attention to the recommendations made. This will help to turn the contents of books and articles into fluent knowledge before their publication. Factors causing text complexity are now studied in the following four groups:

5.1. Presence of a word in the text that has more than one common meaning

Presence of words with several meanings is directly related to text complexity. Depending on their roles in sentences, some words take on special meanings. Sometimes, the presence of a word with several meanings results in several interpretation of the texts. In daily conversations, people naturally understand the meaning of a word with several meanings by paying attention to the topic of the conversation. However, accurate understanding of cases where the presence of a word with several meanings causes ambiguity and complexity is not easy in text processing by computers [4]. Many applied programs have been developed for removing ambiguity from words in written texts. Having insight into the domain of intertextuality substantially helps in automatically recognizing whether a word is ambiguous for the reader or not.

In a single sentence, the meaning allocated to each word is usually determined by the two factors of the syntactic format embedded in the word and the semantics of the words (syntactic dependency) [3]. For example, in sentence (a) the word "absorb" must have a meaning close to "financial affairs" because of its proximity to the word "cost" which is related to finance. On the other hand, in sentence (b) it comes before the word "information" and, therefore, it can be related to a meaning associated with the domain of "learning or of being informed". In sentence (c), the word "treated" is related to the medical field and has the meaning of "being cured," while in sentence (d) the same word is associated with the domain of human relations and has the meaning of "behavior."

a. The customer will *absorb* this *cost*. [Asset] ("pay")
b. The customer will *absorb* this *information*. [Information] ("learn")
c. Peter *treated* Mary with *antibiotics*. [with Medication] ("medical")
d. Peter *treated* Mary with *respect*. [with Quality] ("human relations")

The above sentences illustrate that determination of the domain related to each word can play a major role in removing its ambiguity in the sentence. Of course, in this project we are only after recognizing ambiguity and, therefore, if we find a word with several meanings in a sentence and the domain close to this word is not clear, we need to specify the ambiguity present in the word. The important point to be taken into account at this point is the frequency of the usage of each meaning of the word in the text. That is, if a word has two meanings one of which is rarely used, it should not be included among ambiguous words. However, if a phrase like "as well as" has two

different meanings that are used with relatively equal frequency, it may be ambiguous under certain conditions [7].

5.2. Complexities related to pronouns

Presence of pronouns that are distanced from their antecedents in the text makes sentences ambiguous. For example, if the antecedent of a pronoun comes a few sentences before the pronoun, it will be very difficult for the reader to find the antecedent's exact location. Moreover, sometimes it is not clear in sentences exactly to what an adjective refers [5]. For example, in the sentence "little girls and boys" it is not clear whether the writer meant "all little girls and boys" or "little girls and little boys." Many applied programs have been written for recognizing pronoun antecedents in which accuracies of software performance regarding various kinds of pronouns have been mentioned. The maximum accuracy is that of recognizing the antecedent for the second person pronoun "you," and the minimum those of recognizing the antecedents for the pronouns "it" and "they."

S. Lappin and M. McCord conducted a study only on personal, number, gender pronouns in 1990, and determined a governor for the sentence containing the pronouns [5]. For example, in the sentence "I saw her," the word "saw" is the governor for the words "I" and "her." The relationship between the pronouns in the sentence is then determined depending on the governing literary type. In the section on pronoun recognition, it is impossible to analyze the data with labeling it. Table 2 offers an example of classifying verbs. This classification has been employed in data mining through thematic modeling.

5.3. Presence of several easy synonyms for a word

Simple synonyms and equivalents can be used in place of heavy and difficult ones to simplify a text. For example, replacing "terminate" with "finish" can make understanding the text easier for readers. It is recommended that the text be once matched with the essential database in an English dictionary and, if suitable synonyms are found, they should replace the words in the text. This way words can be included in the text that are familiar to most readers. Databases of essential words for each specific field must be prepared in this stage. In thesis [10], the two sources of everyday literature for children and search engines were used to find the essential words.

Table 2: Thematic modeling

General topics	Examples of verbs
Cognition	Think, analyze, judge…
Communications	Tell, ask, teach …
Feelings	Feel, love, fear …
Social affairs	Participate, make, establish …

5.4. The role played by passive sentences in text ambiguity

If we were to caution the author of an article to omit passive sentences, our recommendations for the text would certainly increase. On the other hand, in some cases the author may not have the necessary information and be forced to use passive sentences. Therefore, and considering the fact that passive sentences are involved in text complexity, the proposed solution is to find the percentage of passive sentences used in each paragraph. If a high percentage of passive sentences are used in the text, the author can be informed to correct the text. Text processing software

recognizes the percentage of passive sentences very accurately. Research has shown passive sentences are most commonly employed in daily conversations but used less frequently, and in descending order, in stories, magazines, newspapers, and university textbooks. This shows omission of passive sentences in university texts must receive special attention in order to remove ambiguities from them.

6. FEASIBILITY OF THE PROPOSED SYSTEM

first point to consider is that it certainly will not be easy to simplify masses of articles (large-volume data). For example, the IEEE Xplore site alone presents more than 20000 articles each year in its research-university databases. Therefore, it is impossible to simplify the entire human knowledge into simple texts. The only thing that can be done is to work case by case. Since we plan to offer a system for making changes in texts to turn them into easy-to-read materials, we will investigate the possibility of implementing some of the processes included in the system that recommends simple reading.

Figures (eg, Figure 1) must be numbered consecutively, 1, 2, etc., from start to finish of the paper, ignoring sections and subsections. Tables (eg, Table 1) are also numbered consecutively, 1, 2, etc., from start to finish of the paper, ignoring sections and subsections, and independently from figures

Table 2: Feasibility of the proposed system

Questions raised in the discussion on the feasibility of the software system	Answering the questions asked about feasibility
Feasibility of implementing the project of simplifying complex texts and of turning them into fluent ones	It may be asked if this is a practical project. Considering the explanations offered in the article, this simplification is carried out in modules and in the different parts of the language; i.e., simplification at the level of words and grammar (in sections such as pronouns, synonyms, words with several meaning, and omission of passive sentences). Therefore, it is possible to add up these modules and create a system for recognizing the effective points causing ambiguity in a text. Of course, this software will be required to recognize 2-3 ambiguities in each page of the text and, certainly, it cannot be expected to find all the complexities in the text due to the inherent ambiguity of languages. In some cases, omission of ambiguities or complexities in a sentence will destroy the whole meaning of the sentence, and all these exceptional cases will be considered.
The scope of the simplifying project	Conference proceedings, various journals, books, and all knowledge containing university texts
Limits of simplifying knowledge containing texts	Considering every university publication has readers of different age with various education levels, an accurate level of simplification cannot be defined. However, simplification must be carried out to a level that the material, while not becoming very boring for readers with mastery over the language, has the difficult expressions simplified as much as possible for those at the beginner level of language proficiency. Moreover, simplification must not include minor points so that the author will have the patience to make the changes. One of the policies in the implementation of this project is for the related knowledge to get close to the level understandable for most people while not falling below university level.

Feasibility of complete implementation of the simplification system proposed in the article just by using the software	Language possesses features that make its analysis and processing very difficult. Our proposed software will be able to discover text complexities as much as possible just by employing data mining and text mining algorithms. Of course, only a human can remove problems with pronouns and only a human can recognize whether the meaning of a sentence changes by replacing a word or by substituting a word with its synonym [11]. Therefore, the software recognizes the probable need for correcting the text, and a human will confirm this recognition and will make the necessary changes. Another example in this relation is the situation in which a person utters a sentence about a subject and the listeners do not understand whether the speaker agrees, overall, with the subject in question or has stated the sentence to express his disagreement. This shows that understanding all the literal complexities and mysteries of a language is difficult, and sometimes impossible, even for humans.
Feasibility of simplifying articles and books for readers with three different levels of language proficiency (the advanced, intermediate, and beginner levels)	An ideal simplification must edit the text for all language proficiency levels. Theoretically, this is partially possible using the conventional method, but its intelligent execution is not possible for software. For example, the book "New Cutting Edge Intermediate" was written for intermediate-level language students. Moreover, other books have been written to teach language to those at other levels of language proficiency. Preparing three easy-to-read versions for each article or book (a very simplified version for the beginner level, a moderately simplified one for the intermediate level, and one in which only the very complex points are simplified for those with an advance level of proficiency in English) will certainly raise the value of the effort made. However, as previously mentioned, humans carry out a great part of the simplification process and, therefore, it will become impossible to prepare three simplified versions for each text because this will require an excessive amount of time. Therefore, the text is simplified only once for all its readers (and this, in itself, has its own complexities).
Feasibility of simplifying previously written knowledge containing sources (articles, books, etc.)	Preparing a database of simple articles with names such as Easy IEEE, easy Amazon, or Easy Springer will greatly help dissemination of sciences. However, as was said above, our proposed project is executed through simplification carried out by the authors only. Considering the possibility of text misinterpretation by a third person leading to a change in the purpose for which the article or book was written, simplification is only carried out by the author and not by a third party. To solve this problem, a formal procedure can be introduced for simplification to be performed by somebody other than the author of the text.
"You do wonders Saadi in speaking so sweetly" Is it possible for the intelligent software system to use stories, examples, or to place suitable images and illustrations in the text?	No. The proposed system is not after eloquent and sweet expression in the text because recognition of questions such as " more explanation needed in the text, or another example must be added" are metaheuristic questions that computerized text processing systems cannot simply understand [12]. Moreover, our system only processes texts and not images or speech; therefore, it will not be possible to include images into texts in our system to simplify texts.
The future of the project	This project requires substantial funds and, therefore, attempts will be made to execute it using the shortest possible algorithms on a small scale. We intend to employ a limited number of algorithms with short execution times instead of a large number with long execution times. The reason for this is that text-processing algorithms naturally require high time complexity. Therefore, considering our limited hardware facilities, we need to use only a small number of fast algorithms. Moreover, we will use small databases instead of large ones. Because of the high processing costs involved, this project will initially be able only to analyze texts paragraph by paragraph. We plan to execute various algorithms on each paragraph in the text, and send the results obtained from each algorithm as our output recommendation. In doing this, we will always refer to the online sites Summarization and Google Translate as successful examples in text processing.

7. CONCLUSIONS

Google has reported that more than 85% of searches made by users have been for finding useless information in the internet space. Moreover, according to the announcement made by the International Publishers Association, only 60% of the books published in Europe in 2013 contained knowledge-enhancing materials, and the rest were attractive novels or were in the entertainment or political domains, etc. The need people have for knowledge, and the fact that people have become comfort seekers, have led to the path toward acquiring useless information that most often lacks knowledge and is used by the public because it is cheap (or free) , rapidly accessible, enjoyable, and easy to read. Therefore, if a project can be executed to turn knowledge-containing complex texts into simple, explicit, expressive, and highly readable ones, it will naturally be possible to take a long step toward making knowledge acquisition easy. In this research, a preliminary project was studied for creating software to change complex texts into simple ones and to remove their ambiguities.

REFERENCES

[1] C. Blake, J. Kampov, A. K. Orphanides, D. West, and C. Lown, "Query expansion, lexical simplification and sentence selection strategies for multi-document summarization," in Document understanding conference (DUC-2007), 2007.

[2] J. Allanson and S. H. Fairclough, "A research agenda for computing," Interacting with Computers, 2004.

[3] J. Medero and M. Ostendorf, "Analysis of vocabulary difficulty using wiktionary," in Proc. SLaTE Workshop, 2009.

[4] Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering polysemic subcategorization frame distributions semantically, pages64–71.

[5] Lappin and M. McCord. 1990. A syntactic filter on pronominal anaphora for slot grammar. In Proceedings of 28th ACL, pages 135–142.

[6] Max, A. (2006). "Writing for language-impaired readers". In Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics, pages 567–570, Mexico City. Springer-Verlag.

[7] Petersen, S. E. and Ostendorf, M. (2007). "Text simplification for language learners: A corpus analysis". In Proceedings of the Speech and Language Technology for Education Workshop (SLaTE-2007), pages 69–72, Pennsylvania, USA.

[8] S. Crossley, D. Allen, and D. McNamara, "Text simplification and comprehensible input: A case for an intuitive approach," Language Teaching Research, vol. 16, no. 1, pp. 89–108, 2012.

[9] S. E. Petersen and M. Ostendorf, "Text simplification for language learners: A corpus analysis," in Speech and Language Technology for Education workshop, 2007.

[10] Siddharthan, A. (2003). Syntactic Simplification and Text Cohesion. PhD thesis, Universityof Cambridge.

[11] V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of ACL/EACL, pages 174–181.

[12] Witten, I., Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann