A Review of Automated Text Simplification

Curtis Bowman

April 22, 2015

Abstract

Text Simplification modifies the syntax and lexicon of a text to improve readability, and understandability of language for the reader. Text simplification has many applications, including: summarization, assistive technologies, and pre-processing for automatic parsing. There are many approaches to the task, including: lexical simplification, syntactic simplification, and statistical machine translation. Text simplification is a difficult task and is a rapidly growing research field. This review gives an overview of the modern research while highlighting the most promising research directions to move the field forward.

1 Introduction

The research topic of automatic text simplification has been steadily growing into it's own field over the past 20 years. Developments in computing power, natural language processing methods, and increased availability of large corpora have allowed steady progress in the field. There are many motivating factors for automatic text simplification. In the United States, 21 to 23 percent of people demonstrated the lowest level of literacy skills. Readers at this level were "apt to experience considerable difficulty in performing tasks that required them to integrate or synthesize information from complex or lengthy texts."[1] To be able to read and fully understand complicated texts is of deep importance.

It is a non-trivial task to supply more readable, simplified texts to struggling readers; this group of people are non-homogeneous and there is no silver bullet for simplifying a text in a way that makes it suitable for all readers. There are a myriad of reasons why people struggle with reading comprehension. Poor readers may suffer from aphasia, dyslexia, or another cognitive

[1]National Center for Education Statistics (http://www.nces.ed.gov).

disability, but this group also includes second language learners, and adults that never had proper reading instruction[7]. Manual text simplification is an incredibly slow process, and is therefore remarkably expensive.

Automatic text simplification aims to reduce the lexical and syntactic complexity of a text, while preserving the meaning and semantics. It aims to create systems that automatically make texts easier to read for a certain target population. Not only can these systems help distribute information to people with learning disabilities, and second language learners, but they can also make texts in technical fields such as healthcare more accessible to the public[5]. There are also many text simplification systems that aim to use it as a pre-processing step for other natural language processing tasks. The idea being that a simplified input will result in better processing and produce better results[4].

2 Approaches

2.1 Lexical Approaches

The most straightforward way to simplify a complex text is to take long or difficult words and replace them with shorter, simpler synonyms. This is the goal of lexical simplification techniques. There is no effort made in these methods to simplify the grammar of a text, instead they are focused on reducing the complexities in vocabulary.

There are traditionally four steps to lexical simplification. Firstly the complex words or phrases in a document must be identified. Secondly, a list of candidate substitutes must be generated for each of these. Thirdly, each set of candidate substitutes must be filtered so that only candidates that make sense in the given context remain. Finally, each set of candidate replacements must be ranked with respect to their complexity. The simplified text is then created by replacing each complex word identified in the input with it's highest ranking candidate replacement [10].

Early attempts at lexical simplification used dictionaries and thesauruses to create a data-set of complex words and their simplest synonyms. This type of data is very limited as it doesn't take into account of language subtleties involving vocabulary. More recent techniques use statistical methods to learn synonym replacements from a data-set. These methods work fairly well, but are limited by the quality of their training data[2, 4, 7].

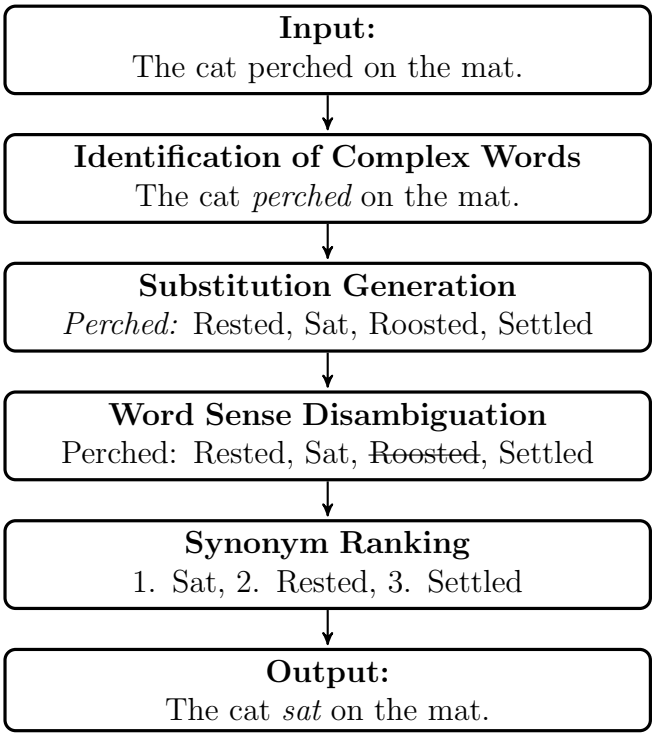

Figure 1: The lexical simplification pipeline.

2.2 Syntactic Approaches

Unlike lexical approaches, syntactic simplification takes the grammatical structure of a sentence into account when trying to simplify it. Sentences with complex grammatical structures lead to increased ambiguity and requires a reader to focus on the structure of the sentence instead of the content. By breaking a long, structurally complex sentence into smaller and simpler sub sentences, the clarity and readability of a text can be improved. Consider the following sentences:

> *The embattled Major government survived a crucial vote on coal pits closure as its last-minute concessions curbed the extent of Tory revolt over an issue that generated unusual heat in the House of Commons and brought the miners to London streets.*

Sentences like the above one are not uncommon when reading complex texts. This can be compared to a simplified(manually) multi-sentence version:

> *The embattled Major government survived a crucial vote on coal pits closure. Its last-minute concessions curbed the extent of Tory*

revolt over the coal-mine issue. The issue generated unusual heat in the House of Commons. It also brought the miners to London streets.[1]

The second version has been split up into smaller sentences, so each sentence then has fewer elements. This reduces the overall complexity of the text, and removes many of the ambiguities from the original.

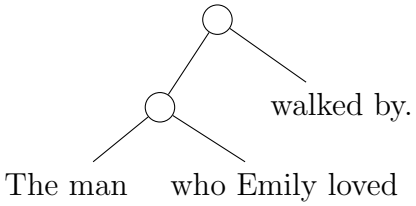

Figure 2: The sentence is analyzed for transformation structures.

This type of simplification is incredibly useful as a pre-processing step for other natural language processing techniques where the clarity of input text is important, such as input machine translation or information retrieval. By creating simpler sub-sentences the text is easier to automatically parse.

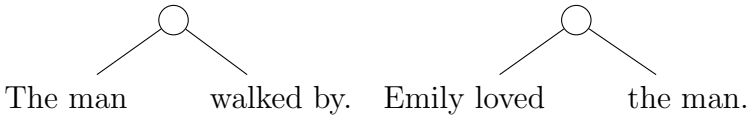

Figure 3: The text is transformed into two simpler sentences that retain meaning.

Syntactic simplification methods have evolved over time, but they began as rule based systems. Hand crafting simplification rules is time-consuming and therefore of limited practical value. Using an aligned text corpus of sentences and their corresponding simplifications, it is possible to induce these simplification rules automatically[1]. However, regardless of how the simplification rules have been created, the steps to then simplify a sentence using these rules is the same. First a sentence is analyzed for transformation structures defined in the set of simplification rules, this can be seen in Figure 2. Then the sentence structure is broken up and transformed using a simplification rule, this can be seen in Figure 3[8, 9].

2.3 Explanation Generation

Explanation generation is a technique of making a complex concept in a text easier to understand by adding supplemental information. While this is not strictly a method of text simplification, there are certain scenarios where it is more successful at increasing the clarity of a document than are other text simplification methods.

Healthcare content is an domain where explanation generation has often proved superior to other simplification methods[5]. Healthcare text are often full of technical terms that are important to preserve for healthcare professionals. However, we also want this content to be accessible to the general public. By using explanation generation we can preserve the technical content of a document while providing supplemental information that makes it appropriate for a general audience as well.

Take the following sentence for example:

"Common causes of myocardial infarction *(heart attack)* include..."

Without the additional information it would be difficult for non-heathcare professionals to understand that this text is about heart attacks. By adding a small annotation to a technical term we can transform a text that was originally created for domain experts, and make it accessible to wider audience of readers.

2.4 Statistical Machine Translation

Automated statistical machine translation is a well established technique in natural language processing. It converts the lexicon and syntax of a source language into that of a target language. English language text simplification can be viewed as a special subset of machine translation where the source language is complex English and the target language is simple English[10].

Machine translation techniques borrow from all of the previous text simplification techniques. By modifying a text with a wide range of operations such a lexical changes, sentence restructures, insertions and deletions, machine translation techniques aim for a generalizeable solution to text simplification.[3, 6]

3 Gaps & Future Research

There are two main hurdles for the text simplification community as a whole:

1. Modern natural language processing systems are data driven and crucially reliant on access to large corpora. Development of new corpora should be a primary concern of the research community.

2. Text simplification systems are difficult to evaluate because of the reliance on hand produced ratings. Progress needs to be made in automatic evaluation systems for text simplification tasks. Without the ability to automate the evaluation of a system, progress will be necessarily slow[11].

There are many small and incremental improvements to be made on specific techniques and methods, however the two issues listed above are the large problems that hold back every form of text simplification. If the text simplification community can address these issues it should propel progress in every area of the field.

4 Conclusion

Text simplification is a growing field that has emerged as a reaction to the need for simplified text. There are many different practical applications for text simplification ranging from helping second language learners, people with cognitive disabilities such as aphasia, and a a pre-processing step in machine translation and other natural language processing tasks.The usefulness of such applications is apparent, and as text simplification systems become more sophisticated text simplification will become a household product, though most individuals will never realize they are using it.

There are multiple approaches to the task. At the lexical level complex words are replaced by simpler ones, at the syntax level the whole structure of a sentence may change in order to simplify it by reducing complex grammatical structures. Then there are systems that utilize machine translation or explanation generation techniques. As the field continues to grow more techniques are certainly to be discovered, and improvements in current systems will be made.

References

[1] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Systems* , vol. 10, no. 3, pp. 183-190, 1997.

[2] M. Yatskar, B. Pang, C. Danescu-Niculescun-Mizil, and L. Lee, "For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 268-272.

[3] W. Coster and D. Kauchak, "Learning to simplify sentences using Wikipedia," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation.* Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 1-9.

[4] O. Biran, S. Brody, and N.Elhadad, "Putting it simply: a context-aware approach to lexical simplification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: shirt paper - Volume 2.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 496-501.

[5] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A semantic and syntactic text simplification tool for health content," in *AMIA Annual symposium Proceedings.* American Medical Informatics Association, 2010, pp. 366-370.

[6] S. Wubben, A. van den Bosch, and E. Krahmer, "Sentence simplification by monolingual machine translation," in *Proceedings of the 50th Annual Meetings of the Association for Computational Linguistics (Volume 1: Long Papers).* Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1015-1024.

[7] R. Keskisärkkä, "Automatic text simplification via synonym replacement." Ph.D. dissertation, Linköping, 2012.

[8] D. Feblowitz and D. Kauchak, "Sentence simplification as tree transduction." in *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations.* Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp 1-10.

[9] S. Klerke and A. Søgaard, "Simple, readable sub-sentences," in 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, 99. 142-149.

[10] M. Shardlow, "A survey of automated text simplification," in *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing.* 2014.

[11] A. Siddharthan, "A survey of research on text simplification," in *International Journal of Applied Linguistics, Special Issue on Recent Advances in Automatic Readability Assessment and Text Simplification.* 2014. vi, 243 pp. 259-298.