Book Review

Automatic Text Simplification

Horacio Saggion
(Universitat Pompeu Fabra)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 37), 2017, xvi+121 pp; paperback, ISBN 978-1-62705-868-1; hardcover, ISBN 9781681732145, $69.95; ebook, ISBN 978-1-62705-869-8, $39.96; doi:10.2200/S00700ED1V01Y201602HLT032

Reviewed by
Xiaojun Wan
Peking University

Automatic text simplification is a special task of text-to-text generation, and it converts a text into another text that is easier to read and understand, while the underlying meaning and information remains the same. A text simplification system usually replaces difficult or unknown phrases with simpler equivalents and transforms long and syntactically complex sentences into shorter and less complex ones. Here is an example from Siddharthan (2006). The first sentence contains two relative clauses and one conjoined verb phrase, and the text below is the simplified version.

- *Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents, which precedes the full purchasing agents report that is due out today and gives an indication of what the full report might hold.*
- *Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents. The Chicago report precedes the full purchasing agents report. The Chicago report gives an indication of what the full report might hold. The full report is due out today.*

Research on automatic text simplification started 20 years ago, and it has become a very important area in natural language processing (NLP) and attracted much attention in recent years. Text simplification facilitates the adaptation of the textual material. It also helps to make texts easier to process and use, thus making accessible information for all a reality. However, it is not an easy task because it involves lexical, syntactic, and semantic issues and possesses interesting challenges.

This book, written by Horacio Saggion, covers all key issues in text simplification, including automatic readability assessment, lexical simplification, syntactic simplification, machine learning–based simplification, and applications of text simplification. Moreover, it describes several typical full text simplification systems, introduces text simplification evaluation techniques, and offers available resources for research and development. This book is written in an elegant way and I enjoyed reading it. It consists of nine chapters, with each chapter focused on a single specific topic about text simplification. Readers can easily choose one or several (almost) self-contained chapters covering the topic(s) they are interested in.

doi:10.1162/coli_r_00332

Chapter 1 briefly introduces what automatic text simplification is and why we need such a technology. It introduces two different text simplification tasks to address different sub-problems: lexical simplification and syntactic simplification. Text simplification tools will be very useful for a large range of users with reading difficulties (e.g., people with aphasia or an autism spectrum disorder). Such tools can be used to create adapted versions of texts for specific populations. I appreciate the text simplification task because it is a promising NLP technology for social good.

Chapter 2 provides an overview of the topic of readability assessment, which is relevant to many approaches to automatic text simplification. Readability assessment techniques can be used to determine the complexity of a given text and thus compare the outputs of different text simplification systems. It is not trivial to automatically assess the readability level of a text. In this chapter, several classical readability formulas are presented and discussed, including the Flesch Reading Ease Score, the Flesch-Kincaid readability formula, the FOG readability score, and the SMOG readability score. Then more robust methods relying on rich syntactic and semantic features are described, along with a few typical classification or regression algorithms (e.g., SVM, kNN, logistic regression) that have been used in these methods.

Chapter 3 covers techniques addressing the lexical simplification problem by replacing words and phrases with simpler equivalents. It introduces the first approach to lexical simplification based on WordNet (Carroll et al. 1998) and then a Spanish lexical simplification system based on word sense disambiguation. After that, approaches based on comparable corpora and distributional lexical semantics are described, followed by a description of a numerical expression simplification system. This chapter also covers relevant evaluation challenges on complex word identification and lexical simplification, which provide benchmark data sets for future research.

Chapter 4 covers techniques to address the syntactic simplification problem by simplifying the syntactic structure of sentences and phrases. Note that this chapter introduces only rule-based approaches, which do not require large annotated corpora. It describes the first syntactic simplification approach targeting constructions such as relative clauses and appositions (Chandrasekar, Doran, and Srinivas 1996) and the approach using typed dependencies, followed by a full rule-based simplification system implemented as a pipeline with four main components. Other approaches relying on information extraction and generation are also outlined. This chapter contains a few examples of rules, which are very helpful for readers' understanding.

Chapter 5 covers techniques to learn simplification from corpora. This chapter first introduces methods that cast text simplification as monolingual machine translation (Specia 2010), and then it surveys methods that use a statistical syntactic-tree translation process. It also surveys optimization techniques for rule application and recent techniques to incorporate semantic information into the simplification problem. With the growth of computing power and data scale, learning-based simplification methods have become prevalent. Neural network models, especially sequence-to-sequence models, have been successfully applied on text simplification. Readers are encouraged to read more recent papers to supplement this chapter.

Chapter 6 presents three fully fledged text simplification systems for different readerships and languages: PSET for English (Carroll et al. 1998), Simplext for Spanish (Saggion et al. 2011), and PorSimples for Brazilian Portuguese (Aluísio and Gasperin 2010). The systems have been designed for specific target populations (people with aphasia, people with low literacy, and people with cognitive disabilities). The details of the systems are described clearly, which will benefit practitioners to build their own text simplification systems.

Chapter 7 introduces various applications of automatic text simplification. This chapter first introduces applications of text simplification for specific target populations (not covered in Chapter 6) and then describes the use of text simplification to facilitate other NLP tasks, including parsing, information extraction, and text summarization. I personally like to see the usefulness of text simplification for other NLP tasks, and I also expect to see a more extensive use of text simplification in the NLP field.

Chapter 8 covers two important topics for text simplification—the available data sets for experimentation and the current evaluation techniques. This chapter introduces lexical resources (English and Non-English corpora for text simplification) that are very useful for building and testing text simplification models or systems. After that, this chapter provides automatic evaluation metrics and methods, including machine learning–based ones. Note that automatic evaluation is not accurate enough and human evaluation is usually more reliable than automatic evaluation for text simplification. How to develop more accurate automatic evaluation metrics or techniques is a research direction in this area.

Chapter 9 concludes, with an overview of the field of text simplification and a critical view of the current state-of-the-art approaches.

In summary, Saggion's book provides a comprehensive and in-depth introduction to automatic text simplification, covering both methodologies and resources for text simplification. It is worth noting that the materials for further reading at the end of each chapter are also valuable resources. This book is recommended not only to researchers, students, and practitioners who work in the area of text simplification, but also to a wide range of readers who are interested in this area. There is still a long way to go in this interesting area, and I look forward to seeing new approaches and resources of text simplification.

References

Aluísio, Sandra Maria and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the Porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, CA.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, Madison, WI.

Chandrasekar, Raman, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics-Volume 2*, pages 1041–1044, Copenhagen.

Saggion, Horacio, Elena Gómez Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in Simplext: Making text more accessible. *Procesamiento del lenguaje natural*, 47:341–342.

Siddharthan, Advaith. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Specia, Lucia. 2010. Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, pages 30–39, Porto Alegre.

Xiaojun Wan is a professor in the Institute of Computer Science and Technology, Peking University, Beijing, China. His research interests are mainly in natural language processing, including text generation, document summarization, sentiment analysis, and semantic computing. His e-mail address is `wanxiaojun@pku.edu.cn`.