Data-Driven Text Simplification

Sanja Štajner
Data and Web Science Group
University of Mannheim, Germany
sanja@informatik.uni-mannheim.de

Horacio Saggion
TALN Research Group
Universitat Pompeu Fabra, Spain
horacio.saggion@upf.edu

Abstract

Automatic text simplification is the process of transforming a complex text into an equivalent version which would be easier to read or understand by a target audience, or easier to handle by automatic natural language processors. The transformation of the text would entail modifications at the vocabulary, syntax, and discourse levels of the text. Over the last years research in automatic text simplification has intensified not only in the number of human languages being addressed but also in the number of techniques being proposed to deal with it from initial rule-based approaches to current data-driven techniques. The aim of this tutorial is to provide a comprehensive overview of past and current research on automatic text simplification.

1 Introduction

Automatic text simplification (ATS) appeared as an area of research in natural language processing (NLP) in the late nineties (Chandrasekar et al., 1996). Its goal is to automatically transform given input (text or sentences) into a simpler variant without significantly changing the input original meaning (Saggion, 2017). What is considered a simpler variant clearly depends on who/what is the target readership/application. Initially, ATS was proposed as a pre-processing step to improve various NLP tasks, e.g. machine translation, information extraction, summarisation, and semantic role labeling. In such scenario, a simpler variant is the one that improves the performance of the targeted NLP task, when used instead of the original input text. Later, the main purpose of ATS systems shifted towards better social inclusion of people with various reading and cognitive impairments, e.g. people with low literacy levels, non-native speakers, people with aphasia, dyslexia, autism, or Down's syndrome. In that case, a simpler variant of a given text snippet would be the one that can be read faster and understood better by the target population.

Traditionally, two different tasks are considered in ATS (Saggion, 2018): lexical simplification is concerned with the modification of the complex or uncommon vocabulary of the text by replacing it with synonyms which are simpler to read or understand, while syntactic simplification is concerned with transforming sentences containing syntactic phenomena which may hinder readability and comprehension (e.g. complex subordination phenomena, passive voice constructions) into simpler equivalents.

Several ATS projects were conducted aimed at producing simplification systems for different audiences and languages. The PSET (Practical Simplification of English Texts) project was a UK initiative to produce adapted texts for aphasic people (Carroll et al., 1998). The PorSimples project (Aluisio et al., 2010) developed an automatic system and editing assistance tool to simplify texts for people with low literacy levels in Brazil. The Simplext project (Saggion et al., 2015) developed simplification technology for Spanish speakers with intellectual disabilities. The FIRST project (Martín-Valdivia et al., 2014) developed a semi-automatic text adaptation tool for English, Spanish and Bulgarian to improve accessibility of written texts to people with autism, while the Able to Include project (Saggion et al., 2017; Ferrés et al., 2016) targeted people with intellectual disabilities on the Web. All those projects had a strong multidisciplinary as well as social character, extending the limits of psycholinguistics, readability assessment, computational linguistics, and natural language processing. We will present the techniques used to transform written texts in each of those projects and make an in-depth discussion of what those projects had in common in terms of techniques and resources, and in what they differed.

1.1 Data-driven Paradigm in Simplification

With the emergence of Simple English Wikipedia and its (comparable) alignment with English Wikipedia, which for the first time offered a large parallel dataset for training of the ATS systems, the approaches to ATS shifted from rule-based (Siddharthan, 2006) to purely data-driven (Coster and Kauchak, 2011; Zhu et al., 2010; Kauchak, 2013), and later hybrid ones (Siddharthan and Mandya, 2014). It created opportunity for stronger NLP component of the systems and new challenges in text/sentence generation, but at the cost of blurring the final goal of those ATS systems, as there was no clear target population in mind anymore. The release of Newsela dataset (Xu et al., 2015) for English and Spanish in 2015, created opportunities for better modelling of simplification operations, given its well-controlled quality of manual simplifications at five different text complexity levels. Following the previously proposed idea of approaching ATS as a monolingual machine translation (MT) task (Specia, 2010; Coster and Kauchak, 2011), Xu et al. (2016) proposed an MT-based ATS system for English built upon Newsela and the large paraphrase database (Pavlick and Callison-Burch, 2016). The manual sentence alignment of English Newsela (Xu et al., 2015), improved automatic alignment of EW-SEW corpus (Hwang et al., 2015), and the recently released free tools for sentence alignment (Paetzold et al., 2017; Štajner et al., 2017; Štajner et al., 2018), offered new opportunities for data-driven ATS.

In 2017, several ATS systems exploring various deep learning architectures appeared, using the new alignments of Wikipedia and Newsela for training. Sequence-to-sequence neural models (Nisioi et al., 2017; Štajner and Nisioi, 2018), and the neural model based on reinforcement learning techniques (Zhang and Lapata, 2017) showed a dominance of neural ATS approaches over the previous data-driven approaches in terms of quality of generated output (better grammaticality and meaning preservation). The question of simplicity of the generated output and the adaptability of those models to different text genres and languages other than English, is still present. While solving the problems of grammaticality and meaning preservation, the neural TS systems introduced a new challenge, showing problems in dealing with abundance of name entities present both in news articles and Wikipedia articles.

2 Tutorial Overview

In this tutorial, we aim to provide an extensive overview of automatic text simplification systems proposed so far, the methods they used and discuss the strengths and shortcomings of each of them, providing direct comparison of their outputs. We aim to break some common misconceptions about what text simplification is and what it is not, and how much it has in common with text summarisation and machine translation. We believe that deeper understanding of initial motivations, and an in-depth analysis of existing TS methods would help researchers new to ATS propose even better systems, bringing fresh ideas from other related NLP areas. We will describe and explain all the most influential methods used for automatic simplification of texts so far, with the emphasis on their strengths and weaknesses noticed in a direct comparison of systems outputs. We will present all the existing resources for TS for various languages, including parallel manually produced TS corpora, comparable automatically aligned TS corpora, paraphrase- and synonym- resources, TS-specific sentence-alignment tools, and several TS evaluation resources. Finally, we will discuss the existing evaluation methodologies for TS, and necessary conditions for using each of them.

3 Tutorial Outline

- Motivation for automatic text simplification:
  - Problems for various NLP tools and applications
  - Reading difficulties of various target populations
- Text simplification projects:
  - Short description of TS projects (PSET, Simplext, PorSimples, FIRST, SIMPATICO, Able to Include)
  - Discussion about the TS projects (what they share and in what they differ)

- Text simplification resources:
  - Resources for lexical simplification
  - Resources for lexico-syntactic simplification
  - Resources for languages other than English
- Evaluation of text simplification systems:
  - Automatic evaluation
  - Human evaluation
- Comparison of non-neural text simplification approaches:
  - Rule-based systems
  - Data-driven systems (supervised and unsupervised)
  - Hybrid systems
  - Semantically-motivated ATS systems
- Neural text simplification (NTS):
  - State-of-the-art neural text simplification (NTS) systems
  - Direct comparison of NTS systems
  - Strengths and weaknesses of NTS systems

4 About the Instructors

Sanja Štajner is currently a postdoctoral research fellow at the University of Mannheim, Germany. She holds a multiple Masters degree in Natural Language Processing and Human Language Technologies (Autonomous University of Barcelona, Spain and University of Wolverhampton, UK) and the PhD degree in Computer Science from the University of Wolverhampton on the topic of "Data-driven Text Simplification". She participated in Simplext and FIRST projects on automatic text simplification, and is the lead author of four ACL papers on text simplification (including the first neural text simplification system) and numerous other papers on the topics of text simplification and readability assessment at various leading international conferences and journals. Sanja's interests in text simplification include building tools for automatic sentence alignment, building ATS systems using various approaches (machine translation, neural machine translation, event-detection, unsupervised lexical simplification), complex word identification (from eye-tracking data, and crowdsourced data), and evaluation of text simplification systems. Sanja regularly teaches NLP at Masters and PhD levels, delivers invited talks and seminars at various universities and companies, and had a very successful tutorial on Deep Learning for Text Simplification at RANLP 2017. She is an area chair for COLING 2018, and regular program committee member of ACL, EMNLP, LREC, IJCAI, IAAA and other international conferences and journals. She was a lead organizer of the first international workshop and shared task on Quality Assessment of Text Simplification (QATS) in 2016, and an organizer of Complex Word Identification shared task in 2018.

Horacio Saggion is an associate professor at the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona. He is the head of the Large Scale Text Understanding Systems Lab associated to the Natural Language Processing group where he works on automatic text summarization, text simplification, information extraction, sentiment analysis and related topics. His research is empirical combining symbolic, pattern-based approaches and statistical and machine learning techniques. Before joining Universitat Pompeu Fabra as a Ramon y Cajal Fellow in 2010, he worked at the University of Sheffield for a number of UK and European research projects developing competitive human language technology. He was also an invited researcher at John Hopkins University for a project on multilingual text summarization. Horacio has been the principal investigator of several national and international projects. Horacio has published over 150 works in leading scientific journals, conferences, and books in the field of human language technology. He organized four international workshops in the areas of text summarization and information extraction and was chair of SEPLN 2014 and co-chair of STIL 2009. He is co-editor of a book on multilingual, multisource information extraction and summarization (Springer 2013) and published a book on Automatic Text Simplification (Morgan & Claypool Publishers 2017). Horacio is a member of the ACL, IEEE, ACM, and SADIO. He is a regular programme committee member for international conferences such as ACL, EACL, COLING, EMNLP, IJCNLP, IJCAI and is an active reviewer for international journals in computer science, information processing, and human language technology. Horacio has international experience in teaching and in addition to his teaching duties at UPF (and previously at Universidad de Buenos Aires) has given intensive courses, tutorials, and invited talks at a number of international events including LREC, ESSLLI, IJCNLP, NLDB, RANLP and RuSSIR.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Carroll, 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 AI and Assistive Technology*, pages 7–10.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *16th International Conference on Computational Linguistics*, pages 1041–1044.

William Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa'ed. 2016. YATS: Yet another text simplifier. In *Proceedings of the 21st International Conference on Applications of Natural Language to Information Systems*, pages 335–342.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*, pages 211–217.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL, pages 1537–1546. The Association for Computer Linguistics.

Maria Teresa Martín-Valdivia, Eugenio Martínez Cámara, Eduard Barbu, Luis Alfonso Ureña López, Paloma Moreda, and Elena Lloret. 2014. Proyecto FIRST (flexible interactive reading support tool): Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos. *Procesamiento del Lenguaje Natural*, 53:143–146.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91.

Gustavo H. Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and Annotation of Comparable Documents. In *The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations*, pages 1–4.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148. ACL.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarević. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Horacio Saggion, Daniel Ferrés, Leen Sevens, Ineke Schuurman, Marta Ripollés, and Olga Rodríguez. 2017. Able to read my mail: An accessible e-mail client with assistive technology. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, pages 5:1–5:4, New York, NY, USA. ACM.

Horacio Saggion. 2017. *Automatic Text Simplification*. Vol.32. Morgan & Claypool Publishers, first edition edition.

Horacio Saggion. 2018. Text simplification. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics (2nd edition)*, chapter 48. Oxford University Press.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 722–731, Gothenburg, Sweden, April. Association for Computational Linguistics.

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

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, PROPOR, pages 30–39, Porto Alegre, RS, Brazil, April 27-30.

Sanja Štajner and Sergiu Nisioi. 2018. A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 97–102.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A Tool for Customized Alignment of Text Simplification Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Associaton for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING, pages 1353–1361, Beijing, China, August.