Natural Language Engineering (2020), **26**, pp. 489–492
doi:10.1017/S1351324919000603

CAMBRIDGE
UNIVERSITY PRESS

Book Review

Horacio Saggion, *Automatic Text Simplification*. Synthesis lectures on human language technologies, April 2017. 137 pages, ISBN:1627058680 9781627058681

Automatic text simplification (TS) is a text-to-text transformation task where the aim is to produce a simpler version of an original text. There are several important aspects to such a task:

- Audience: It is important to know the audience for which the simplified text is intended. Different audiences require different types of simplification operations. For example, the simplification requirements for a second language learner of English may be different from the requirements of a dyslexic person.
- Simplification type: Traditionally, researchers have divided TS in two classes, lexical and syntactic simplification (SS). Lexical simplification (LS) consists of the replacement of words and phrases in the text, while SS encompasses operations at the sentence structure, such as splitting a sentence or changing it from passive to active voice. However, new state-of-the-art approaches deal with both phenomena (machine-translation (MT)-like systems, e.g.).
- Method: As with other areas in Natural Language Processing (NLP), TS can also be explored through rule-based, data-driven or hybrid approaches.

In this book, the author draws largely on his own experiences in order to address the basics of TS and to describe traditional work on this topic. In the introduction chapter, TS is defined alongside explanations of lexical and syntactic types of simplification. This chapter also includes a discussion about *how* texts are simplified by humans and *what* we can then learn in order to automatise the process. Finally, the motivation behind developing TS techniques is presented as a task with social impact that can enable different audiences to access different types of information.

Chapter 2 provides an important overview of work on Readability Assessment (RA), that is, the analysis of textual complexity. Although this topic may not be always linked to TS, knowing the complexity of a text or sentence can be seen as a pre-processing step for TS or as part of its evaluation. Although previous work in TS has highlighted the problems regarding RA shallow metrics (e.g., Shardlow 2014), further development in this area can certainly improve the results and evaluation of TS systems. Perhaps, the biggest contribution in this chapter is Section 2.3, where advanced approaches for RA are discussed.

LS is presented in Chapter 3, which highlights the work of Carroll et al. (1998) as the first approach for LS. The task of complex word identification (CWI) also appears in this chapter, where systems from the SemEval 2016 CWI shared task are detailed (Specia, Jauhar and Mihalcea 2012). The author also highlights that the availability of parallel data from original and simplified texts has promoted the new generation of TS systems, which could then rely on machine learning (ML) approaches. Notably, the most widely used parallel data are the Wikipedia–Simple Wikipedia (W-SW)-aligned data set (e.g., Coster and Kauchak 2011). LS through language modelling (also discussed in this chapter) is another approach only possible due to the acquisition of large amounts of data. Finally, an interesting topic discussed in Chapter 3 is the simplification of numerical expressions in texts. Although this may not be a trendy topic in the LS area, the author highlights the importance of this task for some audiences. In general, this chapter gives good insights about LS, including the author's own work for languages other than English (in this case, Spanish). Paetzold (2016), however, proposes a more structured way of presenting LS and a more complete survey of the topic.

Chapter 4 is dedicated to SS and is a good overview of how rule-based systems work for this task. The discussion about the work of Siddharthan (2006) and Siddharthan (2011) represents a useful survey for researchers wanting a quick reference for the topic. The detailed description of how a rule-based system for SS works, and how it can be used through a well-known toolkit for NLP – General Architecture for Text Engineering (GATE) (Maynard et al. 2002) – makes this chapter even more useful for other researchers.

State-of-the-art approaches for TS are currently data-driven, and Chapter 5 reviews some of the work that has focused on these. TS can be viewed as an MT task, where the original document is the source and the simplified document is the target. Therefore, with the availability of parallel data for TS (such as W-SW), it is natural to explore MT-based approaches for this task (e.g. Zhu, Bernhard and Gurevych 2010; Coster and Kauchak 2011). Another approach explored by Woodsend and Lapata (2011) proposes learning TS rules from parallel data using quasi-synchronous grammars. Finally, Narayan and Gardent (2014) use a hybrid approach that relies on semantic information for split and delete operations, and an MT-based model for modelling paraphrasing.

Although TS is usually motivated using social aspects (e.g., access to information by everyone), very little work has actually resulted in tools and/or has been evaluated by its intended target audiences. Chapter 6 presents three large projects that were successful in at least one of the points above:

- PSET (e.g., Carroll et al. 1998): TS systems to adapt texts for aphasics – English language.
- Simplext (Saggion et al. 2015): TS systems targeting people with intellectual disabilities – Spanish language.
- PorSimples (e.g., Aluísio and Gasperin 2010): TS systems for people with low literacy – Portuguse language.

Chapter 7 continues to present the usefulness of TS either as a tool to help different target audiences (e.g., people with dyslexia; Rello 2014), or as a pre-processing step to improve other NLP tasks (e.g., parsing (Jonnalagadda *et al.* 2009)). Work on TS for specific domains, such as the medical one (Ong et al. 2007), is also discussed in this chapter.

Finally, Chapter 8 is dedicated to presenting resources and tools available for TS, and evaluation approaches for this task. Resources for RA and parallel corpora for TS are described, including a dedicated section on the Simplext data set and a brief discussion about the Newsela data set. The LEXenstein toolkit for LS (Paetzold and Specia 2015) is also presented and discussed. For evaluation, the traditional three-way human evaluation is presented, where human judges are asked to give a *Likert* score (usually from 1 to 5) in order to assess grammaticality, meaning preservation, and simplicity of automatically simplified texts. The author then correctly criticises work that uses RA metrics to evaluate simplified sentences, since such metrics are designed to work on longer texts. Automatic evaluation metrics from the MT area, such as BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002), are also discussed, as these have been largely employed by work on data-driven TS. Unfortunately, the author does not include the work of Xu et al. (2016) that presents SARI – System output Against References and Input sentence – as an automatic metric more adequate for TS. Similar to BLEU, SARI is an n-gram-based metric that takes into account simplification references produced by humans and the original text in order to evaluate the output of a TS system. Data-driven work on TS is now mainly evaluated in terms of SARI, since BLEU is proven to be inadequate (e.g. Sulem, Abend and Rappoport 2018*a*). Finally, the findings of a shared task on quality assessment for TS (QATS) (Stajner, Popovic and Béchara 2016) are presented. The idea in QATS derives from work for quality estimation of MT (e.g., Specia, Scarton and Paetzold 2018), where ML approaches are used to build models using human assessments as

https://doi.org/10.1017/S1351324919000603 Published online by Cambridge University Press

labels. The ultimate goal is to create models able to generalise and automatically predict aspects such as grammaticality, meaning preservation and simplicity for unseen data points.

In summary, this book presents a useful overview of the foundation work on TS. Mainly, Chapters 2, 3 and 4 are great contributions for researchers who are either new to the topic or need to find the best references to the topics discussed. Chapters 6 and 7 also contain important information about TS, its applications and successful projects. Nevertheless, TS has been evolving at a very fast pace since this book was published. Between 2017 and now, over 10 new approaches for TS have been proposed for the English language alone, thanks to the advance of neural deep learning techniques (e.g., Nisioi et al. 2017; Zhang and Lapata 2017; Alva-Manchego et al. 2017; Vu et al. 2018; Jonnalagadda, Luis Jörg, Chitta and Graciela 2009; Sulem, Abend and Rappoport 2018*b*; Scarton and Specia 2018; Zhao et al. 2018; Kriz et al. 2019; Dong et al. 2019; Surya et al. 2019). Additionally, apart from SARI, SAMSA – Simplification Automatic evaluation Measure through Semantic Annotation – (Sulem, Abend and Rappoport 2018*c*) was also proposed as a new metric for TS evaluation that uses semantic information in order to better assess sentence-level operations such as splitting. Therefore, although the reader needs to be aware of the limitations imposed on this book by the fast-growing deep learning movement in NLP and also by the growing interest in TS by the NLP community, this book nevertheless represents a useful reference of traditional work in TS.

Dr. Carolina Scarton
Department of Computer Science, University of Sheffield, Sheffield, UK
Email: c.scarton@sheffield.ac.uk

References

Aluísio S.M. and Gasperin C. (2010). Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pp. 46–53. Association for Computational Linguistics.

Alva-Manchego F., Joachim B., Gustavo P., Carolina S. and Lucia S. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 1: Long Papers). Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 295–305 Asian Federation of Natural Language Processing.

Carroll J., Guido M., Yvonne C., Siobhan D. and John T. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pp. 7–10, Madison, WI.

Coster W. and Kauchak D. (2011). Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation, MTTG '11*, pages 1-9. Portland, OR: ACL.

Dong Y., Zichao L., Mehdi R. and Cheung J. C. K. (2019). EditNTS: an neural programmer-interpreter model for sentences-implification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3393–3402, Florence, Italy. Association for Computational Linguistics

Guo H., Ramakanth P. and Mohit B. (2018). Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, NM: Association for Computational Linguistics, pp. 462–476.

Jonnalagadda S., Luis T., Jörg H., Chitta B., and Graciela G. (2009). Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of NAACL HLT 2009: Short Papers*, pages 177–180, Boulder, CO. Association for Computational Linguistics.

Kriz R.J.S., Marianna A., Carolina Z., Gaurav K., Eleni M. and Chris, C.-B. (2019). Complexity-weighted loss and diverse rerankingfor sentence simplification. In *Proceedings of NAACL-HLT 2019*, pp. 3137–3147, Florence, Italy. Association for Computational Linguistics.

Maynard D., Valentin T., Hamish C., Cristian U., Horacio S., Kalina B. and Yorick W. (2002). Architectural elements of language engineering robustness. *Natural Language Engineering* **8**, 257–274.

Narayan S. and Gardent C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Baltimore, MD: Association for Computational Linguistics, pp. 435–445.

Nisioi S., Sanja Š., Ponzetto S.P. and Dinu, L.P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, pp. 85–91. Association for Computational Linguistics.

https://doi.org/10.1017/S1351324919000603 Published online by Cambridge University Press

Ong E., Damay J., Lojico G., Lu K. and Tarantan D. (2007). Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering* **4**, 37–47.

Paetzold G. and Specia H. (2015). LEXenstein: A Framework for Lexical Simplification *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Beijing, China. Association for Computational Linguistics. pp. 85–90.

Paetzold G. H. (2016). *Lexical Simplification for Non-native English Speakers*. PhD Thesis, The University of Sheffield, Sheffield, UK.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*. Philadelphia, PA: ACL, pp. 311–318.

Rello L.D. (2014) *A Text Accessibility Model for People with Dyslexia*. PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Saggion H., Stajner S., Bott S., Mille S., Rello L. and Drndarevic B. (2015). Making it simplext: implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)* **6**, 14.

Scarton C. and Specia L. (2018). Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 712–718.

Shardlow M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA)* Special Issue on Natural Language Processing 2014, 4(1), 58–70.

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

Siddharthan A. (2011). Text simplification using typed dependencies: a comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation, ENLG '11*. Stroudsburg, PA: Association for Computational Linguistics, pp. 2–11.

Specia L., Jauhar S.K. and Mihalcea R. (2012). SemEval 2012 task 1: English lexical simplification. In *Proceedings of the 1st Joint Conference on Lexical Computational Semantics*, SemEval, pp. 347–355, Montréal, Canada. Association for Computational Linguistics.

Specia L., Scarton C. and Paetzold G.H. (2018). *Quality Estimation of Machine Translation*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Raphael, CA.

Stajner S., Popovic M. and Béchara H. (2016). Quality estimation for text simplification. In *Proceedings of the Workshop and Shared Task on Quality Assessment for Text Simplification (QATS)*. Pororoz, Slovenia.

Sulem E., Abend O. and Rappoport A. (2018a). Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 738–744.

Sulem E., Abend O. and Rappoport A. (2018b). Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 162–173.

Sulem E., Abend O. and Rappoport A. (2018c). Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers). New Orleans, LA: Association for Computational Linguistics, pp. 685–696.

Surya S., Mishra A., Laha A., Jain P. and Sankaranarayanan K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2058–2068, Florence, Italy. Association for Computational Linguistics.

Vu T., Hu B., Munkhdalai T. and Yu H. (2018). Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 2: Short Papers). New Orleans, LA: Association for Computational Linguistics, pp. 79–85.

Woodsend K. and Lapata M. (2011). Learning to simplify sentences with Quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK: Association for Computational Linguistics, pp. 409–420.

Xu W., Napoles C., Pavlick E., Chen Q. and Callison-Burch C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* **4**, 401–415.

Zhang X. and Lapata M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 595–605.

Zhao S., Meng R., He D., Andi S. and Bambang P. (2018). Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

Zhu Z., Bernhard D. and Gurevych I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*, Stroudsburg, PA: Association for Computational Linguistics, pp. 1353–1361.

https://doi.org/10.1017/S1351324919000603 Published online by Cambridge University Press