Expert Systems With Applications 82 (2017) 383–395

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Leveraging event-based semantics for automated text simplification

Sanja Štajner*, Goran Glavaš

Data and Web Science Group, University of Mannheim, B6 26, Mannheim DE-68161, Germany

ARTICLE INFO

Article history:
Received 23 November 2016
Revised 31 March 2017
Accepted 1 April 2017
Available online 6 April 2017

Keywords:
Automated text simplification
Event extraction
Semantics

ABSTRACT

Automated Text Simplification (ATS) aims to transform complex texts into their simpler variants which are easier to understand to wider audiences and easier to process with natural language processing (NLP) tools. While simplification can be applied on lexical, syntactic, and discourse level, all previously proposed ATS systems only operated on the first two levels, thus failing at simplifying texts on the discourse level. We present a semantically-motivated ATS system which is the first system that is applied on the discourse level. By exploiting the state-of-the-art event extraction system, it is the first ATS system able to eliminate large portions of irrelevant information from texts, by maintaining only those parts of the original text that belong to factual event mentions. A few handcrafted rules ensure that the output of the system is syntactically simple, by placing each factual event mention in a separate short sentence, while the state-of-the-art unsupervised lexical simplification module, based on using word embeddings, replaces complex and infrequent words with their simpler variants. We perform a thorough evaluation, both automatic and manual, showing that our system produces more readable and simpler texts than the state-of-the-art ATS systems. Our newly proposed post-editing evaluation further reveals that our system requires less human effort for correcting grammaticality and meaning preservation on news articles than the state-of-the-art ATS system.

1. Introduction

Many texts we encounter daily are written using very complex syntactic structures and specialised or sophisticated vocabulary and thus cannot be understood by many readers. Many initiatives raised awareness about this issue, offering guidelines for writing in an easy-to-read manner in order to produce texts more accessible for everyone, including non-native speakers and people with any kind of language or intellectual impairment, e.g. “Make it Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability” (Freyhoff, Hess, Kerr, Tronbacke, & Van Der Veken, 1998) or “Federal Plain Language Guidelines” (PlainLanguage, 2011). However, manual simplification of existing texts is time-consuming and requires very specific training. At the same time, it has been noticed that syntactically complex sentences, as well as infrequent words and phrases, decrease the performance of various NLP tasks, such as parsing (Chandrasekar, Doran, & Srinivas, 1996), machine translation (Chandrasekar, 1994; Štajner & Popović, 2016), information extraction (Beigman Klebanov, Knight, & Marcu, 2004; Evans, 2011), or semantic role labelling (Vickrey & Koller, 2008). For both these reasons, late nineties yielded a new NLP task of Automated Text Simplification (ATS) that aims to (semi-)automatically transform complex texts into their simpler variants that are more understandable to wider audiences and easier to process with various NLP tools.

1.1. Motivation for text simplification

It has been shown that complex sentence structures (passive constructions, long sentences, appositions, etc.) and infrequent or long words can be difficult to understand for many people, e.g. non-native speakers (Petersen & Ostendorf, 2007), people with low literacy levels (Aluísio, Specia, Pardo, Maziero, & Fortes, 2008), and people with different kinds of reading or cognitive impairments, such as dyslexia (Rello, 2012), aphasia (Devlin & Unthank, 2006), autism spectrum disorders (Martos, Freire, González, Gil, & Sebastian, 2012), or Down’s syndrome (Saggion et al., 2015).

The writing style in newspaper articles is particularly challenging. They often contain long sequences of adjectives, e.g.*“twenty-five-year-old blond-haired mother-of-two Jane Smith”* (Carroll, Minnen, Canning, Devlin, & Tait, 1998), which can cause problems for people with aphasia (Carroll et al., 1998), autism spectrum

* Corresponding author.
E-mail addresses: sanja@informatik.uni-mannheim.de (S. Štajner), goran@informatik.uni-mannheim.de (G. Glavaš).

http://dx.doi.org/10.1016/j.eswa.2017.04.005
0957-4174/

disorders (Martos et al., 2012), and intellectual disabilities (Feng, 2009). In order to present the information in a more sensational way, the newspaper articles often use passive constructions which do not follow the canonical subject–verb–object structure and thus pose difficulties to people with aphasia (Carroll et al., 1999) or autism spectrum disorders (Martos et al., 2012). For example, instead of using the straightforward active voice which follows the canonical subject–verb–object structure *"The council today accepted a bid to build an incinerator on local wasteland"*, it is more common to find the same information in a passive sentence *"A bid to build an incinerator on local wasteland was today accepted by the council"* (Carroll et al., 1998).

At the word level, it has been noticed that infrequent words can be difficult to understand for people with aphasia (Devlin, 1999) and autism spectrum disorders (Martos et al., 2012; Norbury, 2005) and lead to a longer reading time in people with dyslexia (Rello, Baeza-Yates, Dempere-Marco, & Saggion, 2013).

At the discourse level, people with autism spectrum disorders or intellectual disabilities may have problem finding main idea and inferring information (Feng, 2009; Martos et al., 2012), resolving anaphors (Ehrlich, Remond, & Tardieu, 1999; Martos et al., 2012; Shapiro & Milkes, 2004) and understanding texts written in dialog format (Drndarević & Saggion, 2012; Martos et al., 2012). Furthermore, long texts pose additional problems to people with intellectual disabilities, as they have difficulties processing and retaining large amounts of information (Fajardo et al., 2014; Feng, 2009) and suppressing irrelevant information (Gernsbacher & Faust, 1991). They can also affect self-efficacy and reading motivation in students with intellectual disability (Gómez, 2011; Morgan & Moni, 2008).

Long and syntactically or semantically complex sentences are not only difficult to understand for humans, but they can also pose difficulties to machine processing. Many studies have thus tried to (manually) simplify such sentences in a pre-processing step in order to improve the performance of various NLP tools. It has been noticed that simple sentences generate a smaller number of possible parse trees and have fewer constituents which leads to a faster and less ambiguous parsing (Chandrasekar et al., 1996). Such sentences are also easier to process by machine translation systems due to a simpler sentence structure, simpler vocabulary and less ambiguity (Chandrasekar, 1994; Štajner & Popović, 2016). Vickrey and Koller (2008) showed that a rule-based sentence simplification system used as a pre-processing step significantly improves results of the semantic role labelling (SRL) task. Beigman Klebanov et al. (2004) showed that the use of Easy Access Sentences (EAS) – the sentences with only one tensed verb and in which pronouns are substituted with the appropriate names – lead to better performance of information retrieval systems. The use of simplified sentences also improves information extraction in medical texts (Evans, 2011).

1.2. Current problems in automated text simplification

So far, the majority of the proposed ATS systems were built for English, ranging from the early-days rule-based systems (Carroll et al., 1998; Devlin, 1999; Siddharthan, 2006), through data-driven approaches based on the comparable English Wikipedia – Simple English Wikipedia (EW–SEW) corpus (Coster & Kauchak, 2011b) using phrase-based statistical machine translation (Coster & Kauchak, 2011a; Kauchak, 2013; Štajner, Bechara, & Saggion, 2015a; Wubben, van den Bosch, & Krahmer, 2012) or syntactic machine translation (Woodsend and Lapata, 2011a; Zhu, Berndard, & Gurevych, 2010), and more recent hybrid approach (Siddharthan & Angrosh, 2014) that combines supervised data-driven lexical simplification with rule-based syntactic simplification.

All rule-based syntactic simplification modules proposed so far require a significant amount of handcrafted rules. For example, the system proposed by Siddharthan and Angrosh (2014) contains 26 hand-crafted rules for apposition, relative clauses, and combinations of the two; 85 rules for subordination and coordination, 11 for conversion from passive to active voice, and 14 for the standardisation of quotations. The lexical simplification modules are usually supervised and require parallel dataset for training, which limits them to the EW–SEW corpus (approx. 160,000 sentence pairs) and thus reduces their coverage. Our systems, in contrast, do not require a large number of handcrafted rules for syntactic simplification module (they only use 11 rules in total) and our lexical simplification module is fully unsupervised, thus not requiring any parallel or comparable text simplification datasets for training.

Additional problem with most of the existing ATS systems is that they do not perform any kind of content reduction, while at the same time, they make simplified texts often longer than the original texts by performing sentence splitting and adding explanations of difficult terms. Long texts – although lexically and syntactically simpler – can again pose problems to people with intellectual disabilities as they have problem with memory load and with suppressing irrelevant information (see Section 1.1). The analysis of manual simplifications of texts for people with intellectual disabilities, done by trained human editors familiar with the user needs and following specific guidelines, revealed that human editors often delete irrelevant information, some sentence parts or even whole sentences (Drndarević, Štajner, Bott, Bautista, & Saggion, 2013; Petersen & Ostendorf, 2007). However, apart from the three ATS systems (Angrosh, Nomoto, & Siddharthan, 2014; Narayan & Gardent, 2014; Woodsend & Lapata, 2011a) which perform some very light content reduction (occasionally delete an adjective phrase or a sentence argument), there have been no ATS systems which would address this important issue. Unlike those, our system performs transformations not only on lexical and syntactic levels but also on the discourse level, leading thus to significantly more content reduction within a sentence and within a text (deleting even whole sentences).

1.3. The goal and contributions

We propose an end-to-end ATS system that overcomes all aforementioned shortcomings and dedicates special attention to content reduction.[1] The event-based simplification (EBS) module is based on the state-of-the-art event extraction system (Glavaš & Šnajder, 2015) and uses only 11 rules to perform sentence splitting and deletion of irrelevant sentences or sentence parts. The lexical simplification (LS) module leverages word embeddings trained on a large (standard English) corpora, thus not requiring any parallel or comparable TS corpora. Yet, it performs comparably well as, or better than, the state-of-the-art supervised lexical simplification model proposed by Horn, Manduca, and Kauchak (2014).

We also examine how the order of the simplification modules influences the system performance. On one hand, the lexical substitutions performed on the original text can influence event detection system (if we first apply the LS module) and thus lead to different selection of sentences to be retained as relevant. Applying the EBS module before the LS module, on the other hand, can lead to increased number of correct or incorrect substitutions, due to the repetition of the event actors during the sentence splitting process. Therefore, we perform an in-depth manual error analysis on 475 sentences simplified by two different system configurations: LexEv (first applying the LS module and then the EBS module), and EvLex (first applying the EBS module and then the LS module).

[1] This article builds upon and expands on Glavaš and Štajner (2015; 2013).