Sentence Simplification for Semantic Role Labelling and Information Extraction

Richard Evans
Research Institute in Information and Language Processing
University of Wolverhampton
United Kingdom
r.j.evans@wlv.ac.uk

Constantin Orăsan
Research Institute in Information and Language Processing
University of Wolverhampton
United Kingdom
c.orasan@wlv.ac.uk

Abstract

In this paper, we report on the extrinsic evaluation of an automatic sentence simplification method with respect to two NLP tasks: semantic role labelling (SRL) and information extraction (IE). The paper begins with our observation of challenges in the intrinsic evaluation of sentence simplification systems, which motivates the use of extrinsic evaluation of these systems with respect to other NLP tasks. We describe the two NLP systems and the test data used in the extrinsic evaluation, and present arguments and evidence motivating the integration of a sentence simplification step as a means of improving the accuracy of these systems. Our evaluation reveals that their performance is improved by the simplification step: the SRL system is better able to assign semantic roles to the majority of the arguments of verbs and the IE system is better able to identify fillers for all IE template slots.

1 Introduction

Sentence simplification is one aspect of text simplification, which is concerned with the conversion of texts into a more accessible form. In many cases, text simplification is performed to facilitate subsequent human or machine text processing. This may include processing for human reading comprehension (Canning, 2002; Scarton et al., 2017; Orăsan et al., 2018) or for NLP tasks such as dependency parsing (Jelínek, 2014), information extraction (Jonnalagadda et al., 2009; Evans, 2011; Peng et al., 2012), semantic role labelling (Vickrey and Koller, 2008), and multidocument summarisation (Blake et al., 2007; Siddharthan et al., 2004).

In previous research, Caplan and Waters (1999) noted a correlation between sentence comprehension difficulty for human readers and the numbers of propositions expressed in the sentences being read.[1] Evans and Orăsan (2019) presented an iterative rule-based approach to sentence simplification which is intended to reduce the per sentence propositional density of input texts by converting sentences which contain compound clauses and complex NPs[2] into sequences of simpler sentences.

Evaluation of text simplification systems is difficult, especially when such evaluations need to be conducted repeatedly for development purposes and cost is a critical factor. In general, the choice of evaluation method depends on the purpose of the simplification task. Various types of evaluation are currently used, but these are problematic. In previous work, evaluation of sentence simplification systems (including Evans and Orăsan's (2019) system, which is extrinsically evaluated in our current paper) has relied on one or more of three main approaches: the use of overlap metrics such as Levenshtein distance (Levenshtein, 1966), BLEU score (Papineni et al., 2002) and SARI (Xu et al., 2016) to compare system output with human simplified texts (e.g. Wubben et al., 2012; Glavas and Stajner, 2013; Vu et al., 2014); automated assessments of the readability of system output (Wubben et al., 2012; Glavas and Stajner, 2013; Vu et al., 2014); and surveys of human opinions about the grammaticality, readability, and meanings of system output (Angrosh et al., 2014; Wubben et al., 2012; Feblowitz and Kauchak, 2013). In previous work, researchers have also used methods such as

[1] Propositions are atomic statements that express simple factual claims (Jay, 2003). They are considered the basic units involved in the understanding and retention of text (Kintsch and Welsch, 1991).

[2] NPs which contain finite nominally bound relative clauses.

Proceedings of Recent Advances in Natural Language Processing, pages 285–294,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_033

eye tracking (Klerke et al., 2015; Timm, 2018), and reading comprehension testing (Orăsan et al., 2018) to evaluate text simplification systems.

There are several challenges in these approaches to evaluation. The development of gold standards in text simplification is problematic because they are difficult to produce and numerous variant simplifications are acceptable. As a result, existing metrics may not accurately reflect the usefulness of the simplification system being evaluated. Even when there are detailed guidelines for the simplification task, there is still likely to be a variety of means by which a human might simplify a text to produce a reference simplification. Further, due to the difficulty of the human simplification task, it may be that evaluation measures such as BLEU and SARI are unable to exploit a sufficiently large set of reference simplifications.

Evaluation of text simplification methods using automatic readability metrics is problematic because the extent to which all but a handful of readability metrics correlate with human reading comprehension is uncertain. Evaluation via opinion surveys of readers is difficult because participants may have varying expectations about the upper and lower limits of sentence complexity, making responses to Likert items unreliable. Participants also vary in terms of linguistic ability and personal background knowledge. These variables, which affect reading behaviour and may affect responses to opinion surveys, are difficult to control.

When using methods such as eye tracking to evaluate text simplification, previous work has shown that differences in reading behaviour depend on participants' reading goals (Yeari et al., 2015). This variable is usually controlled by asking participants to respond to text-related opinion surveys or multiple choice reading comprehension questions. One adverse effect of this is that these evaluations may be of limited validity when considering the usefulness of system output for other purposes. While we may learn whether a sentence simplification method improves participants' performance in answering short reading comprehension questions, it is not clear whether similar benefits would be obtained in terms of readers' abilities to be entertained by the text or to understand it well enough to be able to summarise it for friends.

Given that text simplification is usually made for a particular purpose, the evaluation method should offer insights into the suitability of the text simplification system for this purpose. Extrinsic evaluation offers the possibility of meeting this requirement. Text simplification has also been claimed to improve automatic text processing (e.g. Vickrey and Koller, 2008; Evans, 2011; Hasler et al., 2017), though the evidence for this has been fairly limited. In this paper, we explore whether syntactic simplification can facilitate two NLP tasks: semantic role labelling (SRL) and information extraction (IE).

In Section 2 of this paper, we present an overview of previous related work. In Section 3, we present an overview of Evans and Orăsan's (2019) method for sentence simplification, which is the simplification method used in our current paper. In Section 4, we present each of the extrinsic evaluation experiments based on SRL (Section 4.1) and IE (Section 4.2). Each of these sections describes the task, the test data used, the NLP system whose output is used for extrinsic evaluation of the sentence simplification system, our motivation for considering that accuracy of the NLP system may be improved via a preprocessing step in which sentence simplification is performed, the evaluation method, our results, and a discussion of the results. In Section 5, we draw conclusions and consider directions for future work.

2 Related Work

Chandrasekar and Srinivas (1997) hypothesised that approaches to sentence simplification may evoke improvements in subsequent text processing tasks. In previous work, researchers have sought to determine whether or not a preprocessing step based on text simplification can facilitate subsequent natural language processing. In the current paper, our concern is to investigate the impact of a system simplifying sentences which contain compound clauses. Hogan (2007) and Collins (1999) observed that, for dependency parsers, dependencies involving coordination are identified with by far the worst accuracy of any dependency type (F_1-*score* $\approx 61\%$). This is one factor motivating our research in this direction.

Sentence simplification has also been applied as a preprocessing step in neural machine translation and hierarchical machine translation (Hasler et al., 2017). In their approach, the approach to sentence simplification was sentence compression. One contribution of our current paper is an investigation of the use of an information preserving ap-

proach to sentence simplification as a preprocessing step in the NLP applications.

Vickrey and Koller (2008) applied their sentence simplification method to improve performance on the CoNLL-2005 shared task on SRL.[3] For sentence simplification, their method exploits full syntactic parsing with a set of 154 parse tree transformations and a machine learning component to determine which transformation operations to apply to an input sentence. They find that a SRL system based on a syntactic analysis of automatically simplified versions of input sentences outperforms a strong baseline. In their evaluation, Vickrey and Koller (2008) focus on the overall performance of their SRL system rather than on the particular contribution made by the sentence simplification method. As noted earlier, in our current paper, we isolate sentence simplification as a preprocessing step and investigate its impact on subsequent NLP tasks.

3 Sentence Simplification System

Evans and Orăsan (2019) presented an iterative rule-based method for sentence simplification based on a shallow syntactic analysis step. Their system transforms input sentences containing compound clauses and complex NPs into sequences of simpler sentences that do not contain these types of syntactic complexity.

The first stage of sentence simplification is a shallow syntactic analysis step which tags textual markers of syntactic complexity, referred to as *signs*, with information about the syntactic constituents that they coordinate or of which they are boundaries. The signs of syntactic complexity are a set of conjunctions, complementisers, wh-words, punctuation marks, and bigrams consisting of a punctuation mark followed by a lexical sign. In the analysis step, syntactic constituents are not identified. It is only the signs which are tagged. The automatic sign tagger was developed by Dornescu et al. (2013). In their scheme, clause coordinators are tagged CEV[4] while the left boundaries of subordinate clauses are tagged SSEV.[5]

After shallow syntactic analysis of the sentence, an iterative algorithm is applied to sentences containing compound clauses and complex NPs. The algorithm (Algorithm 1) integrates a sentence transformation function which implements the transformation schemes listed in Table 1.

Input: Sentence s_0, containing at least one sign of syntactic complexity of class c, where $c \in \{\text{CEV, SSEV}\}$.
Output: The set of sentences A derived from s_0, that have reduced propositional density.

1 The empty stack W;
2 $O \leftarrow \emptyset$;
3 $push(s_0, W)$;
4 **while** $isEmpty(W)$ *is* $false$ **do**
5 $pop(s_i, W)$;
6 **if** s_i *contains a sign of syntactic complexity of class* c *(specified in Input)* **then**
7 $s_{i_1}, s_{i_2} \leftarrow transform_c(s_i)$;
8 $push(s_{i_1}, W)$;
9 $push(s_{i_2}, W)$;
10 **else**
11 $O \leftarrow O \cup \{s_i\}$
12 **end**
13 **end**

Algorithm 1: Sentence simplification algorithm

In its original implementation, the *transform* function (line 7 of Algorithm 1) included 28 sentence simplification rules to implement one transformation scheme simplifying compound clauses and 125 rules implementing three transformation schemes simplifying sentences which contain complex NPs. Evaluation of the method revealed that simplification of sentences containing complex NPs was significantly less reliable than simplification of sentences containing compound clauses. For this reason, in the extrinsic evaluations presented in this paper, we deactivated the rules simplifying sentences that contain complex NPs. Each of the remaining implemented rules includes a rule activation pattern which, when detected in the input sentence, triggers an associated transformation operation. Table 1 presents the transformation scheme used to simplify compound clauses and an example of the sentence transformation that it makes. Input sentences are transformed if they match any of the rule activation patterns, which are expressed in terms of particular words, parts of speech, and tagged signs of syntactic complexity. Each application of a rule transforms a single input sentence into two sim-

[3] http://www.lsi.upc.edu/~srlconll/spec.html. Last accessed 14th May 2019.
[4] Coordinator of Extended projections of a Verb.
[5] Start of Subordinate Extended projection of a Verb.

Scheme	Input Sentence	Output Sentence 1	Output Sentence 2
A [B –**CEV** C] D. → A B D. A C D.	{They were formally found not guilty by the recorder Michael Gibbon QC after}$_A$ [a witness, who cannot be identified, withdrew from giving evidence$_B$ and**CEV** prosecutor Susan Ferrier offered no further evidence$_C$]{}$_D$.	{They were formally found not guilty by the recorder Michael Gibbon QC after}$_A$ a witness, who cannot be identified, withdrew from giving evidence$_B$ {}$_D$.	{They were formally found not guilty by the recorder Michael Gibbon QC after}$_A$ prosecutor Susan Ferrier offered no further evidence$_C$ {}$_D$

Table 1: *Sentence transformation scheme used to simplify sentences containing compound clauses*

pler sentences which are added to the working set (stack W in Algorithm 1).

The iterative nature of the algorithm enables it to convert complex sentences containing multiple signs of syntactic complexity such as (1) into the sequence of simple sentences (2).

(1) Kattab, of Eccles, Greater Manchester, was required to use diluted chloroform water in the remedy, but the pharmacy only kept concentrated chloroform, which is 20 times stronger.

(2) a. Kattab, of Eccles, Greater Manchester, was required to use diluted chloroform water in the remedy.
b. The pharmacy only kept concentrated chloroform.
c. Concentrated chloroform is 20 times stronger.

4 Experimental Setup

We evaluated the sentence simplification method extrinsically via two NLP applications. In each case, the application was treated as a black box. We compared performance of the system when processing input in its original form and in an automatically simplified form generated by the simplification method. As noted in Section 3, our approach to sentence simplification is syntactic rather than lexical. As they are based to some extent on exact string matching, the experiments described in this paper would be unsuitable for evaluation of lexical simplification systems.

4.1 Semantic Role Labelling

Semantic role labelling (SRL) is the task of automatically detecting the different arguments of predicates expressed in input sentences. We evaluated a system performing SRL in accordance with the Propbank formalism. In this scheme, an "individual verb's semantic arguments are numbered, beginning with zero. For a particular verb, [A0] is generally the argument exhibiting features of a Prototypical Agent (Dowty, 1991), while [A1] is a Prototypical Patient or Theme. No consistent generalizations can be made across verbs for the higher-numbered arguments"[6] (Palmer et al., 2005). The scheme includes semantic roles for "general, adjunct-like arguments" providing information on the verb's cause [AMCAU], direction [AMDIR], discourse relations [AMDIS], location [AMLOC], manner [AMMNR], modal function[7] [AMMOD], negation [AMNEG], purpose [AMPNC], and time [AMTMP], among others. For extrinsic evaluation of the sentence simplification method, we focused on verbal predicates[8], their arguments, and the nine listed adjunct-like argument types.

Table 2 provides an example of SRL to analyse sentence (3).

(3) When Disney offered to pay Mr. Steinberg a premium for his shares, the New York investor didn't demand the company also pay a premium to other shareholders.

The table contains a row of information about the semantic roles associated with each of the four main verbs occurring in the sentence. For example, it encodes information about the agent (*the New York investor*), patient or theme (*the company also pay a premium to other shareholders*), time (*When Disney offered to pay Mr. Steinberg a premium for his shares*), and negation (*n't*) of the verb *demand*.

Test Data. No suitable test data exist to evaluate a SRL system as a means of extrinsically evaluating the sentence simplification method. Although annotated data from the CONLL-2004/5[9] shared tasks on SRL are available, this test data is available only for the original versions of input sentences and not for simplified versions which may be generated using sentence simplification systems. Given that it is difficult to map verbs,

[6] Such as [A2], etc.

[7] Applicable to verbs.

[8] As opposed to prepositional, adjectival, or other types of predicate.

[9] http://www.lsi.upc.edu/~srlconll/home.html. Last accessed 23rd May 2019.

A0	V	A1	A2	A3	AMDIS	AMNEG	AMTMP
Disney	offered	to pay Mr. Steinberg a premium for his shares					
Disney	pay	his shares	Mr. Steinberg	a premium			
the New York investor	demand	the company also pay a premium to other shareholders				n't	When Disney offered to pay Mr. Steinberg a premium for his shares
the company	pay		other shareholders	a premium	also		

Table 2: *Example of semantic role labelling of Sentence (3)*

their arguments, and the semantic labels of these arguments from sentences in their original form to groups of sentences in their automatically generated simplifications, we developed a new set of test data for this purpose. We used a 7270-token collection of news articles from the METER corpus (Gaizauskas et al., 2001) to derive a new manually annotated data set. The original version of this dataset contains 265 sentences while the automatically simplified one contains 470 sentences.

NLP System. We made our extrinsic evaluation of the sentence simplification method using *Senna* (Collobert et al., 2011), a SRL system which tags predicates and their arguments in accordance with the formalism used in *Propbank*

Motivation. In our previous work (Evans and Orăsan, 2019), we used six metrics to assess the readability of the original and simplified versions of texts which include those that we use as test data for the SRL task. We found that the automatically simplified news texts have a lower propositional density (0.483 vs. 0.505) and reading grade level (5.4 vs. 10.3) and greater syntactic simplicity (89.07 vs. 46.81) and temporal consistency, assessed in terms of tense and aspect (30.15 vs. 27.76) than the original news texts. We determined the scores for these readability metrics using the CPIDR tool (Covington, 2012)[10] and the Coh-Matrix Web Tool (McNamara et al., 2014). As a task dependent on accurate syntactic parsing, we would expect that automatic SRL would be more accurate when processing the simplified versions of the input texts.

Evaluation Method. We applied Senna to the original and automatically simplified versions of the test data. Table 3 contains an example of the semantic roles labelled in one of the test sentences that we used. In this table, arguments identified more accurately in simplified sentences are underlined. For cases in which the SRL performed by Senna differed when processing the original and automatically simplified versions of input sentences, we manually inspected the two analyses, and recorded the number of cases for which SRL of the original sentence was superior to that of the simplified sentence, and vice versa. The inspection was made by a single annotator. In future work, we will seek to employ additional annotators for this task.

Results. Our manual evaluation of output from Senna revealed that 86.39% (1707) of the arguments identified in the two versions of the texts were identical. Of the remaining arguments, 5.31% (105) of those correctly identified in the original versions of the texts were not identified in the simplified versions, while 8.29% (164) of the arguments correctly identified in the simplified versions of the texts were not identified in the original versions. Of the 269 arguments identified in only one of the versions of the texts, 60.97% were arguments identified more accurately in the simplified version, while 39.03% were arguments identified more accurately in the original versions of the texts.

Table 4 shows the number of semantic roles labelled more accurately, by type, when Senna processes the original (*Orig*) and the automatically simplified (*Simp*) versions of news articles. To illustrate, when processing the original versions of the news texts, Senna correctly identifies the agents (arguments with semantic role label A0) of 14 verbs that it did not identify when process-

[10] http://ai1.ai.uga.edu/caspr/CPIDR-3.2.zip. Last accessed 31st May 2019.

Original sentence: But Smith had already been arrested - her clothing had been found near his home and DNA tests linked him to it.						
A0	**V**	**A1**	**A2**	**AMDIS**	**AMLOC**	**AMTMP**
	arrested	Smith		But		already
	found	her clothing			near his home and DNA tests linked him to it	
his home and DNA tests	linked	him	to it			
Simplified sentence: But Smith has already been arrested - her clothing had been found near his home. DNA tests linked him to it.						
A0	**V**	**A1**	**A2**	**AMDIS**	**AMLOC**	**AMTMP**
	arrested	Smith		But		already
	found	her clothing			near his home	
DNA tests	linked	him	to it			

Table 3: *Example of more accurate semantic role labelling in automatically simplified text.*

Role	Orig vs. Simp	Simp vs. Orig
A0 (agent)	14	23
A1 (patient/theme)	45	77
A2 (less prominent than A1)	14	13
AMCAU (cause)	0	1
AMDIR (direction)	4	0
AMDIS (discourse relation)	0	3
AMLOC (location)	3	13
AMMNR (manner)	4	6
AMNEG (negation)	0	1
AMPNC (purpose)	1	6
AMTMP (time)	12	27
V (verb)	2	3
Total	**99**	**173**

Table 4: *Positive differences in numbers of true positives obtained for semantic role labelling of original and simplified versions of input texts*

ing the automatically simplified versions of those texts. Conversely, when processing the automatically simplified versions, Senna correctly identified the agents of 23 verbs that it did not identify when processing the original versions.

Discussion. Overall, while there are advantages to performing SRL on each version of input texts, the greatest improvement in performance arises from processing the automatically simplified versions. A larger-scale evaluation is necessary but this observation constitutes some evidence that the sentence simplification method facilitates the NLP task of SRL.

4.2 Information Extraction

Information extraction (IE) is the automatic identification of selected types of entities, relations, or events in free text (Grishman, 2005). In this paper, we are concerned with IE from vignettes which provide brief clinical descriptions of hypothetical patients.

The discourse structure of these vignettes consists of six elements: *basic information* (patient's gender, profession, ethnicity, and health status); *chief complaint* (the main concern motivating the patient to seek medical intervention); *history* (a narrative description of the patient's social, family, and medical history); *vital signs* (a description of the patient's pulse and respiration rates, blood pressure, and temperature); *physical examination* (a narrative description of clinical findings observed in the patient); and *diagnostic study and laboratory study* (the results of several different types of clinical test carried out on the patient).

Each element in the discourse structure is represented by a template encoding related information. For example, the template for physical examinations holds information on each clinical finding/symptom (FINDING) observed in the examination, information on the technique used to elicit that finding (TECHNIQUE), the bodily location to which the technique was applied (LOCATION), the body system that the finding pertains to (SYSTEM),

and any qualifying information about the finding (QUALIFIER). In this article, we focus on automatic extraction of information pertaining to physical examinations. The goal of the IE system is to identify the phrases used in the clinical vignette that denote findings and related concepts and add them to its database entry for the vignette.

Test Data. Our test data comprises a set of 286 clinical vignettes and completed IE templates, encoding information about TECHNIQUEs, LOCATIONs, SYSTEMs, and QUALIFIERs, associated with the 719 FINDINGs that they contain. This test data was developed in the context of an earlier project and is based on clinical vignettes owned by the National Board of Medical Examiners.[11]

NLP System. For the experiments described in this paper, we used a simple IE system in which input texts are tokenised and part of speech tagged, domain-specific gazetteers are used to identify references to medical concepts and a simple set of finite state transducers (FSTs) is used to group adjacent references to concepts into multiword terms. The gazetteers and FSTs were developed in previous work presented by Evans (2011).

After tagging references to clinical concepts in the vignettes, IE is performed using a small number of simple rules. To summarize, vignettes are processed by considering each sentence in turn. Every mention of a clinical FINDING or SYMPTOM is taken as the basis for a new IE template. The first tagged TECHNIQUE, SYSTEM, and LOCATION within the sentence containing the focal SYMPTOM or FINDING is considered to be related to it.[12] QUALIFIERS (e.g. *bilateral* or *peripheral*) are extracted in the same way, except in sentences containing the word *no*. In these cases, the QUALIFIER related to the FINDING is identified as *none*.

The sentences in the test data were simplified using the method presented in Section 3. We then ran the IE system in two settings. In the first (IE_{ORIG}), it processed the original collection of vignettes. In the second (IE_{SIMP}), it processed the automatically simplified vignettes which contain a reduced number of compound clauses.

[11] https://www.nbme.org/. Last accessed 31st May 2019.

[12] Versions of the system in which the closest tagged concept was extracted in each case, rather than the first, were significantly less accurate in both cases (overall accuracy of 0.6542 for IE from the original vignettes, and 0.6567 for IE from vignettes automatically simplified using the system described in Section 3). See Table 5 for results obtained using the superior IE system.

Motivation. An analysis of the readability of the original and simplified versions of the clinical vignettes did not provide a strong indication that the automatic sentence simplification method would improve the accuracy of the IE system. The 286 original clinical vignettes in the test data have a mean propositional density of 0.4826 ideas per word and 5.499 ideas per sentence. The values of these metrics for the simplified versions of the vignettes are 0.4803 ideas per word and 5.269 ideas per sentence, respectively. Although they are of the correct polarity, these differences are not statistically significant ($p = 0.5327$ and $p = 0.1407$, respectively). However, previous work in sentence simplification for IE (Jonnalagadda et al., 2009; Evans, 2011; Peng et al., 2012; Niklaus et al., 2016) has demonstrated that automatic sentence simplification can improve the accuracy of IE systems. This motivated us to evaluate the impact of the automatic sentence simplification method in this task.

Evaluation Method. For the IE task, our evaluation metric is based on F_1-score averaged over all slots in the IE templates and all templates in the test data. Identification of true positives is based on exact matching of system-identified slot fillers with those in the manually completed IE templates in our test data.

Results. The accuracy scores obtained by each variant of the IE system are presented in Table 5. Inspection of this table reveals that FINDINGs and all related concepts are identified more accurately in the simplified versions of the input texts.

Sentence (4) and its automatically simplified variant (5) provide an example of the difference in performance obtained by the two systems. In these examples, identified FINDINGs are italicised and associated concepts are underlined. Multiword terms appear in square brackets.

(4) She has truncal$_{LOC}$ *obesity* and pigmented$_{QUAL}$ abdominal$_{LOC}$ *striae*.

(5) a. She has truncal$_{LOC}$ [*obesity striae*].
b. She has pigmented$_{QUAL}$ abdominal$_{LOC}$ *striae*.

In (5-a), the FINDING *obesity* is not tagged correctly because the SYMPTOM *striae* is erroneously grouped with *obesity* to form a new FINDING, *obesity striae* which does not match the FINDING listed in the gold standard. By contrast, LOCATIONS in (5) are identified with greater accu-

	IE_{ORIG}		IE_{SIMP}		
Template slot	Acc	95% CI	Acc	95% CI	Best Performer
FINDING	0.8819	[0.847, 0.914]	0.8861	[0.853, 0.917]	0.5486
TECHNIQUE	0.8514	[0.814, 0.886]	0.8903	[0.858, 0.922]	0.9344
SYSTEM	0.8097	[0.769, 0.850]	0.8431	[0.806, 0.881]	0.873
QUALIFIER	0.7431	[0.697, 0.786]	0.7708	[0.728, 0.814]	0.794
LOCATION	0.8431	[0.806, 0.881]	0.8611	[0.825, 0.894]	0.735
All	0.8258	[0.808, 0.843]	0.8503	[0.834, 0.867]	0.976

Table 5: *Performance of the IE systems processing our test data.*

racy than those in (4) because IE_{ORIG} erroneously extracts the same LOCATION (*truncal*) for both FINDINGS in (4).

We applied a bootstrapping method to obtain confidence intervals for accuracy of extraction of each of the IE template slots. For this purpose, 50% of the the output of each system was randomly sampled in each of 100 000 evaluations. The confidence intervals are presented in the *95% CI* columns of Table 5. The figures in the *Best Performer* column of this table indicate the proportion of evaluations for which the IE_{SIMP} system was more accurate than the IE_{ORIG} system. Differences in the accuracy of IE were found to be statistically significant in all cases, using McNemar's test ($p < 0.00078$), with the exception of differences when extracting FINDINGS ($p = 0.6766$).

Discussion. Chinchor (1992) notes that assessment of the statistical significance of differences in accuracy between different IE systems is challenging. In our evaluation experiment, Dos Santos et al. (2018) framed the comparison between IE_{ORIG} and IE_{SIMP} using a binomial regression model. Given that such models apply only when the variables being considered are independent, dos Santos et al. (2018) included a latent variable in the analysis to represent the effect of the text on the performance of the two systems (the two evaluations are not independent because both systems process the same text). They showed that the odds ratio of agreement between IE_{SIMP} and the gold standard is 1.5 times greater than that between IE_{ORIG} and the gold standard. For all slots in the IE template, the probability of agreement between IE_{ORIG} and the gold standard is 0.937. The probability of agreement between IE_{SIMP} and the gold standard is 0.957. This difference is statistically significant. They conclude that IE_{ORIG} and IE_{SIMP} differ in their performance on the information extraction task. The probability of agreement with our gold standard is greater for IE_{SIMP} than for IE_{ORIG}, although the probability of agreement is already large for IE_{ORIG}. This evaluation indicates that the automatic sentence simplification method facilitates IE.

5 Conclusions

As a result of various difficulties identified in current approaches to intrinsic evaluation of sentence simplification methods, we performed an extrinsic evaluation of one information-preserving sentence simplification method via three NLP tasks. We found that the sentence simplification step brings improvements to the performance of IE and SRL systems. In a third experiment, not described here due to space restrictions, we evaluated the sentence simplification method extrinsically with respect to a multidocument summarisation task using MEAD (Radev et al., 2006) to summarise clusters of documents developed for Task 2 of DUC-2004.[13] We found that the simplification step had no impact on this task. As a result, although the findings reported in our current paper seem promising, it is difficult to know the extent to which they are applicable to other NLP tasks or to tasks which differ only with respect to the test data used. This is one issue that we are interested in exploring in future work. Another is a test of whether extrinsic evaluation methods sensitive to information about the types of changes made in the simplification step would perform better than the black box methods used in the current paper.

[13]Information about the DUC conferences is accessible from https://www-nlpir.nist.gov/projects/duc/index.html (last accessed 22nd August 2018). Guidelines about the tasks presented in DUC-2004 are available at https://www-nlpir.nist.gov/projects/duc/guidelines/2004.html (last accessed 22nd August 2018).

References

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Association for Computational Linguistics, Dublin, Ireland, pages 1996–2006.

Catherine Blake, Julia Kampov, Andreas K. Orphanides, David West, and Cory Lown. 2007. Uncch at duc 2007: Query expansion, lexical simplification and sentence selection strategies for multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC-2007)*. National Institute of Standards and Technology.

Yvonne Canning. 2002. *Syntactic Simplification of Text*. Ph.d. thesis, University of Sunderland.

David Caplan and Gloria S. Waters. 1999. Verbal working memory and sentence comprehension. *Behavioural and Brain Sciences* 22:77–126.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems* 10:183–190.

Nancy Chinchor. 1992. The statistical significance of the muc-4 results. In *Proceedings of the Fourth Message Understanding Conference*. McLean, Virginia, pages 30–50.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.d thesis, University of Pennsylvania.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537. http://dl.acm.org/citation.cfm?id=1953048.2078186.

Michael A. Covington. 2012. CPIDR® 5.1 user manual. Technical report, Institute for Artificial Intelligence, University of Georgia, Athens, Georgia, U.S.A.

Iustin Dornescu, Richard Evans, and Constantin Orasan. 2013. A Tagging Approach to Identify Complex Constituents for Text Simplification. In *Proceedings of Recent Advances in Natural Language Processing*. Hissar, Bulgaria, pages 221 – 229.

Larissa Sayuri Futino Castro dos Santos, Marcos Oliveira Prates, Gisele de Oliveira Maia, Guilherme Lucas Moreira Dias Almeida, Daysemara maria Cotta, Ricardo Cunha Pedroso, and Aurélio de Aquino araújo. 2018. Assessing if an automated method for identifying features in texts is better than another: discussions and results. Technical report, Department of Statistics, Universidade Federal de Minas Gerais. https://bit.ly/2xUD2BI.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Richard Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing* 26 (4):371–388.

Richard Evans and Constantin Orăsan. 2019. Identifying signs of syntactic complexity for rule-based sentence simplification. *Natural Language Engineering* 25 (1):69–119.

Dan Feblowitz and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1–10.

Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott Piao. 2001. The meter corpus: A corpus for analysing journalistic text reuse. In *Proceedings of Corpus Linguistics 2001 Conference*. Lancaster University Centre for Computer Corpus Research on Language, pages 214–223.

Goran Glavas and Sanja Stajner. 2013. Event-centered simplication of news stories. In *Proceedings of the Student Workshop held in conjunctuion with RANLP-2013*. RANLP, Hissar, Bulgaria, pages 71–78.

Ralph Grishman. 2005. The Oxford Handbook of Computational Linguistics. Oxford University Press, chapter Information Extraction, pages 545–559.

Eva Hasler, Adrià de Gispert, Felix Stahlberg, and Aurelien Waite. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & language* 45:221–235.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 680–687.

Timothy B. Jay. 2003. *The psychology of language*. Pearson, Upper Saddle Rive, NJ.

Tomáš Jelínek. 2014. Improvements to dependency parsing using automatic simplification of data. In *Proceedings of LREC-2014 the 22nd International Conference on Computational Linguistics (Coling 2008)*. European Language Resources Association, Reykjavik, Iceland, pages 73–77.

Siddhartha Jonnalagadda, Luis Tari, Jorg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of NAACL HLT 2009: Short Papers*. Associ-

ation for Computational Linguistics, Boulder, Colorado, pages 177–180.

Walter Kintsch and David M. Welsch. 1991. The construction–integration model: A framework for studying memory for text. In W. E. Hockley and S. Lewandowsky, editors, *Relating theory and data: Essays on human memory*, Hillsdale, NJ: Erlbaum, pages 367–385.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *NODALIDA*. Linköping University Electronic Press / ACL, pages 97–105.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions and Insertions and Reversals. *Soviet Physics Doklady* 10 (8):707–710.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Osaka, Japan, pages 170–174. https://www.aclweb.org/anthology/C16-2036.

Constantin Orăsan, Richard Evans, and Ruslan Mitkov. 2018. Intelligent text processing to help readers with autism. In Khaled Shaalan, Aboul Ella Hassanien, and Mohammed F. Tolba, editors, *Intelligent Natural Language Processing: Trends and Applications*, Springer, pages 713–740.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106. https://doi.org/10.1162/0891201053630264.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, pages 311–318.

Yifan Peng, Catalina O. Tudor, Manabu Torii, Cathy H. Wu, and K. Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, Philadelphia, PA, pages 1–6.

Dragomir Radev, John Blitzer, Adam Winkel, Tim Allison, and Michael Topper. 2006. MEAD documentation v3.10. Technical report, University of Michigan.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A Multilingual Syntactic Simplification Tool. In *The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations*. Taipei, Taiwan, pages 25–28.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '04. https://doi.org/10.3115/1220355.1220484.

Linnea B. Timm. 2018. *Looking at text simplification - Using eye tracking to evaluate the readability of automatically simplified sentences*. Bachelor thesis, Institutionen fr datavetenskap, Linkpings universitet.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*. Columbus, Ohio, pages 344–352.

Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. 2014. *Learning to Simplify Children Stories with Limited Data*, Springer, Bangkok, Thailand, pages 31–41.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12)*. Association for Computational Linguistics, Jeju, Republic of South Korea, pages 1015–1024.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *TACL* 4:401–415. https://bit.ly/2Sj5mag.

Menahem Yeari, Paul van den Broek, and Marja Oudega. 2015. Processing and memory of central versus peripheral information as a function of reading goals: evidence from eye-movements. *Reading and Writing* 28 (8):1071–1097.