Semantic Structural Evaluation for Text Simplification

Elior Sulem, Omri Abend, Ari Rappoport
Department of Computer Science, The Hebrew University of Jerusalem
{eliors|oabend|arir}@cs.huji.ac.il

Abstract

Current measures for evaluating text simplification systems focus on evaluating lexical text aspects, neglecting its structural aspects. In this paper we propose the first measure to address structural aspects of text simplification, called SAMSA. It leverages recent advances in semantic parsing to assess simplification quality by decomposing the input based on its semantic structure and comparing it to the output. SAMSA provides a reference-less automatic evaluation procedure, avoiding the problems that reference-based methods face due to the vast space of valid simplifications for a given sentence. Our human evaluation experiments show both SAMSA's substantial correlation with human judgments, as well as the deficiency of existing reference-based measures in evaluating structural simplification.[1]

1 Introduction

Text simplification (*TS*) addresses the translation of an input sentence into one or more simpler sentences. It is a useful preprocessing step for several NLP tasks, such as machine translation (Chandrasekar et al., 1996; Mishra et al., 2014) and relation extraction (Niklaus et al., 2016), and has also been shown useful in the development of reading aids, e.g., for people with dyslexia (Rello et al., 2013) or non-native speakers (Siddharthan, 2002).

The task has attracted much attention in the past decade (Zhu et al., 2010; Woodsend and Lapata, 2011; Wubben et al., 2012; Siddharthan and Angrosh, 2014; Narayan and Gardent, 2014), but has yet to converge on an evaluation protocol that yields comparable results across different methods and strongly correlates with human judgments. This is in part due to the difficulty to combine the effects of different simplification operations (e.g., deletion, splitting and substitution). Xu et al. (2016) has recently made considerable progress towards that goal, and proposed to tackle it both by using an improved reference-based measure, named SARI, and by increasing the number of references. However, their research focused on lexical, rather than structural simplification, which provides a complementary view of TS quality as this paper will show.

This paper focuses on the evaluation of the structural aspects of the task. We introduce the semantic measure SAMSA (Simplification Automatic evaluation Measure through Semantic Annotation), the first structure-aware measure for TS in general, and the first to use semantic structure in this context in particular. SAMSA stipulates that an optimal split of the input is one where each predicate-argument structure is assigned its own sentence, and measures to what extent this assertion holds for the input-output pair in question, by using semantic structure. SAMSA focuses on the core semantic components of the sentence, and is tolerant towards the deletion of other units.[2]

For example, SAMSA will assign a high score to the output split "John got home. John gave Mary a call." for the input sentence "John got home and gave Mary a call.", as it splits each of its predicate-argument structures to a different sentence. Splits that alter predicate-argument relations such as "John got home and gave. Mary called." are penalized by SAMSA.

SAMSA's use of semantic structures for TS evaluation has several motivations. First, it provides means to measure the extent to which the meaning of the source is preserved in the out-

[1] All data and code are available in https://github.com/eliorsulem/SAMSA.

[2] We do not consider other structural operations, such as passive to active transformations (Canning, 2002), that are currently not treated by corpus-based simplification systems.

put. Second, it provides means for measuring whether the input sentence was split to semantic units of the right granularity. Third, defining a semantic measure that does not require references avoids the difficulties incurred by their non-uniqueness, and the difficulty in collecting high quality references, as reported by Xu et al. (2015) and by Narayan and Gardent (2014) with respect to the Parallel Wikipedia Corpus (PWKP; Zhu et al., 2010). SAMSA is further motivated by its use of semantic annotation only on the source side, which allows to evaluate multiple systems using same source-side annotation, and avoids the need to parse system outputs, which can be garbled.

In this paper we use the UCCA scheme for defining semantic structure (Abend and Rappoport, 2013). UCCA has been shown to be preserved remarkably well across translations (Sulem et al., 2015) and has also been successfully used for machine translation evaluation (Birch et al., 2016) (Section 2). We note, however, that SAMSA can be adapted to work with any semantic scheme that captures predicate-argument relations, such as AMR (Banarescu et al., 2013) or Discourse Representation Structures (Kamp, 1981), as used by Narayan and Gardent (2014).

We experiment with SAMSA both where semantic annotation is carried out manually, and where it is carried out by a parser. See Section 4. We conduct human rating experiments and compare the resulting system rankings with those predicted by SAMSA. We find that SAMSA's rankings obtain high correlations with human rankings, and compare favorably to existing reference-based measures for TS. Moreover, our results show that existing measures, which mainly target lexical simplification, are ill-suited to predict human judgments where structural simplification is involved. Finally, we apply SAMSA to the dataset of the QATS shared task on simplification evaluation (Štajner et al., 2016). We find that SAMSA obtains comparative correlation with human judgments on the task, despite operating in a more restricted setting, as it does not use human ratings as training data and focuses only on structural aspects of simplicity. Section 2 presents previous work. Section 3 discusses UCCA. Section 4 presents SAMSA. Section 5 details the collection of human judgments. Our experimental setup for comparing our human and automatic rankings is given in Section 6, and results are given in Section 7, showing superior results for SAMSA. A discussion on the results is presented in Section 8. Section 9 presents experiments with SAMSA on the QATS evaluation benchmark.

2 Related Work

Evaluation Metrics for Text Simplification. As pointed out by Xu et al. (2016), many of the existing measures for TS evaluation do not generalize across systems, because they fail to capture the combined effects of the different simplification operations. The two main directions pursued are direct human judgments and automatic measures borrowed from machine translation (MT) evaluation. Human judgments generally include grammaticality (or fluency), meaning preservation (or adequacy) and simplicity. Human evaluation is usually carried out with a small number of sentences (18 to 20), randomly selected from the test set (Wubben et al., 2012; Narayan and Gardent, 2014, 2016).

The most commonly used automatic measure for TS is BLEU (Papineni et al., 2002). Using 20 source sentences from the PWKP test corpus with 5 simplified sentences for each of them, Wubben et al. (2012) investigated the correlation of BLEU with human evaluation, reporting positive correlation for simplicity, but no correlation for adequacy. Štajner et al. (2014) explored the correlation with human judgments of six automatic metrics: cosine similarity with a bag-of-words representation, METEOR (Denkowski and Lavie, 2011), TERp (Snover et al., 2009), TINE (Rios et al., 2011) and two sub-components of TINE: T-BLEU (a variant of BLEU which uses lower n-grams when no 4-grams are found) and SRL (based on semantic role labeling). Using 280 pairs of a source sentence and a simplified output with only structural modifications, they found positive correlations for all the metrics except TERp with respect to meaning preservation and positive albeit lower correlations for METEOR, T-BLEU and TINE with respect to grammaticality. Human simplicity judgments were not considered in this experiment. In this paper we collect human judgments for grammaticality, meaning preservation and *structural* simplicity. To our knowledge, this is the first work to target structural simplicity evaluation, and it does so both through elicitation of human judgments and

through the definition of SAMSA.

Xu et al. (2016) were the first to propose two evaluation measures tailored for simplification, focusing on lexical simplification. The first metric is FKBLEU, a combination of iBLEU (Sun and Zhou, 2012), originally proposed for evaluating paraphrase generation by comparing the output both to the reference and to the input, and of the Flesch-Kincaid Index (FK), a measure of the readability of the text (Kincaid et al., 1975). The second one is SARI (System output Against References and against the Input sentence) which compares the n-grams of the system output with those of the input and the human references, separately evaluating the quality of words that are added, deleted and kept by the systems. They found that FKBLEU and even more so SARI correlate better with human simplicity judgments than BLEU. On the other hand, BLEU (with multiple references) outperforms the other metrics on the dimensions of grammaticality and meaning preservation.

As the Parallel Wikipedia Corpus (PWKP), usually used in simplification research, has been shown to contain a large portion of problematic simplifications (Xu et al., 2015; Hwang et al., 2015), Xu et al. (2016) further proposed to use multiple references (instead of a single reference) in the evaluation measures. SAMSA addresses this issue by directly comparing the input and the output of the simplification system, without requiring manually curated references.

Structural Measures for Text-to-text Generation. Other than measuring the number of splits (Narayan and Gardent, 2014, 2016), which only assesses the frequency of this operation and not its quality, no structural measures were previously proposed for the evaluation of structural simplification. The need for such a measure is pressing, given recent interest in structural simplification, e.g., in the Split and Rephrase task (Narayan et al., 2017), which focuses on sentence splitting.

In the task of sentence compression, which is similar to simplification in that they both involve deletion and paraphrasing, Clarke and Lapata (2006) showed that a metric that uses syntactic dependencies better correlates with human evaluation than a metric based on surface sub-strings. Toutanova et al. (2016) found that structure-aware metrics obtain higher correlation with human evaluation over bigram-based metrics, in particular with grammaticality judgments, but that they do not significantly outperform bigram-based metrics on any parameter. Both Clarke and Lapata (2006) and Toutanova et al. (2016) use reference-based metrics that use syntactic structure on both the output and the references. SAMSA on the other hand uses linguistic annotation only on the source side, with semantic structures instead of syntactic ones.

Semantic structures were used in MT evaluation, for example in the MEANT metric (Lo et al., 2012), which compares the output and the reference sentences, both annotated using SRL (Semantic Role Labeling). Lo et al. (2014) proposes the XMEANT variant, which compares the SRL structures of the source and output (without using references). As some frequent constructions like nominal argument structures are not addressed by the SRL annotation, Birch et al. (2016) proposed HUME, a human evaluation metric based on UCCA, using the semantic annotation only on the source side when comparing it to the output. We differ from HUME in proposing an automatic metric, tackling monolingual text simplification, rather than MT.

The UCCA annotation has also been recently used for the evaluation of Grammatical Error Correction (GEC). The USIM metric (Choshen and Abend, 2018) measures the semantic faithfulness of the output to the source by comparing their respective UCCA graphs.

Semantic Structures in Text Simplification. In most of the work investigating the structural operations involved in text simplification, both in rule-based systems (Siddharthan and Angrosh, 2014) and in statistical systems (Zhu et al., 2010; Woodsend and Lapata, 2011), the structures that were considered were syntactic. Narayan and Gardent (2014, 2016) proposed to use semantic structures in the simplification model, in particular in order to avoid splits and deletions which are inconsistent with the semantic structures. SAMSA identifies such incoherent splits, e.g., a split of a phrase describing a single event, and penalizes them.

Glavas and Štajner (2013) presented two simplification systems based on event extraction. One of them, named Event-wise Simplification, transforms each factual event motion into a separate sentence. This approach fits with SAMSA's stipulation, that an optimal structural simplification is one where each (UCCA-) event in the input

sentence is assigned a separate output sentence. However, unlike in their model, SAMSA stipulates that not only should multiple events evoked by a verb in the same sentence be avoided in a simplification, but penalizes sentences containing multiple events evoked by a lexical item of any category. For example, the sentence "John's unexpected kick towards the gate saved the game" which has two events, one evoked by "kick" (a noun) and another by "saving" (a verb) can be converted to "John kicked the ball towards the gate. It saved the game."

3 UCCA's Semantic Structures

In this section we will briefly describe the UCCA scheme, focusing on the concepts of Scenes and Centers which are key in the definition of SAMSA. UCCA (Universal Cognitive Conceptual Annotation; Abend and Rappoport, 2013) is a semantic annotation scheme based on typological (Dixon, 2010b,a, 2012) and cognitive (Langacker, 2008) theories which aims to represent the main semantic phenomena in the text, abstracting away from syntactic detail. UCCA structures are directed acyclic graphs whose nodes (or units) correspond either to the leaves of the graph (including the words of the text) or to several elements jointly viewed as a single entity according to some semantic or cognitive consideration. Unlike AMR, UCCA semantic units are directly anchored in the text (Abend and Rappoport, 2017; Birch et al., 2016), which allows easy inclusion of a word-to-word alignment in the metric model (Section 4).

UCCA Scenes. A Scene, which is the most basic notion of the foundational layer of UCCA considered here, describes a movement, an action or a state which persists in time. Every Scene contains one main relation, which can be either a Process or a State. The Scene may contain one or more Participants, which are interpreted in a broad sense, including locations and destinations. For example, the sentence "He ran into the park" has a single Scene whose Process is "ran". The two Participants are "He" and "into the park".

Scenes can have several roles in the text. First, they can provide additional information about an established entity (Elaborator Scenes) as for example the Scene "who entered the house" in the sentence "The man who entered the house is John". They can also be one of the Participants of another Scene, for example, "he will be late" in the sentence: "He said he will be late". In the other cases, the Scenes are annotated as parallel Scenes (H) which can be linked by a Linker (L): "When$_L$ [he will arrive at home]$_H$, [he will call them]$_H$".

Unit Centers. With regard to units which are not Scenes, the category Center denotes the semantic head of the unit. For example, "dogs" is the center of the expression "big brown dogs" and "box" is the center of "in the box". There could be more than one Center in a non-Scene unit, for example in the case of coordination, where all conjuncts are Centers.

4 The SAMSA Metric

SAMSA's main premise is that a structurally correct simplification is one where: (1) each sentence contains a single event from the input (UCCA Scene), (2) the main relation of each of the events and their participants are retained in the output.

For example, consider "John wrote a book. I read that book." as a simplification of "I read the book that John wrote.". Each output sentence contains one Scene, which has the same Scene elements as the source, and would thus be deemed correct by SAMSA. On the other hand, the output "John wrote. I read the book." is an incorrect split of that sentence, since a participant of the "writing" Scene, namely "the book" is absent in the split sentence. SAMSA would indeed penalize such a case.

Similarly, Scenes which have elements across several sentences receive a zero score by SAMSA. As an example, consider the sentence "The combination of new weapons and tactics marks this battle as the end of chivalry", and erroneous split "The combination of new weapons and tactics. It is the end of chivalry." (adapted from the output of a recent system on the PWKP corpus), which does not preserve the original meaning.

4.1 Matching Scenes to Sentences

SAMSA is based on two external linguistic resources. One is a semantic annotation (UCCA in our experiments) of the source side, which can be carried out either manually or automatically, using the TUPA parser[3] (Transition-based UCCA parser; Hershcovich et al., 2017) for UCCA. UCCA decomposes each sentence s into a set of Scenes $\{sc_1, sc_2, .., sc_n\}$, where each scene

[3] https://github.com/danielhers/tupa

sc_i contains a main relation mr_i (sub-span of sc_i) and a set of zero or more participants A_i.

The second resource is a word-to-word alignment A between the words in the input and one or zero words in the output. The monolingual alignment thus permits SAMSA not to penalize outputs that involve lexical substitutions (e.g., "commence" might be aligned with "start"). We denote by n_{inp} the number of UCCA Scenes in the input sentence and by n_{out} the number of sentences in the output.

Given an input sentence's UCCA Scenes $sc_1, \ldots, sc_{n_{inp}}$, a non-annotated output of a simplification system split into sentences $s_1, \ldots, s_{n_{out}}$, and their word alignment A, we distinguish between two cases:

1. $n_{inp} \geq n_{out}$: in this case, we compute the maximal Many-to-1 correspondence between Scenes and sentences. A Scene is matched to a sentence in the following way. We say that a leaf l in a Scene sc is *consistent* in a Scene-sentence mapping M which maps sc to a sentence s, if there is a word $w \in s$ which l aligns to (according to the word alignment A). The score of matching a Scene sc to a sentence s is then defined to be the total number of consistent leaves in sc. We traverse the Scenes in their order of occurrence in the text, selecting for each the sentence that maximizes the score. If $n_{inp} = n_{out}$, once a sentence is matched to a Scene, it cannot be matched to another one. Ties between sentences are broken towards the sentence that appeared first in the output.

$$\mathrm{M}^*(sc_i) = \mathrm{argmax}_s score(sc_i, s)$$
$$\text{s.t. } s \notin \{M^*(sc_1), \ldots, M^*(sc_{i-1})\} \text{ if } n_{inp} = n_{out}$$

2. $n_{inp} < n_{out}$: In this case, a Scene will necessarily be split across several sentences. As this is an undesired result, we assign this instance a score of zero.

4.2 Score Computation

Minimal Centers. The minimal center of a UCCA unit u is UCCA's notion of a semantic head word, defined through recursive rules not unlike the head propagation rules used for converting constituency structures to dependency structures. More formally, we define the minimal center of a UCCA unit u (here a Participant or a Main Relation) to be the UCCA graph's leaf reached by starting from u and iteratively selecting the child tagged as Center. If a Participant (or a Center inside a Participant) is a Scene, its center is the main relation (Process or State) of the Scene.

For example, the center of the unit "The previous president of the commission" (u_1) is "president of the commission". The center of the latter is "president", which is a leaf in the graph. So the minimal center of u_1 is "president".

Given the input sentence Scenes $\{sc_1, ..., sc_{n_{inp}}\}$, the output sentences $\{s_1, ..., s_{n_{out}}\}$, and a mapping between them M^*, SAMSA is defined as:

$$\frac{n_{out}}{n_{inp}} \frac{1}{2n_{inp}} \sum_{sc_i} [\mathbb{1}_{M^*(sc_i)}(MR_i) + \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbb{1}_{M^*(sc_i)}(\mathrm{Par}_i^{(j)})]$$

where $\mathrm{MR_i}$ is the minimal center of the main relation (Process or State) of sc_i, and $\mathrm{Par_i^{(j)}}$ ($j = 1, \ldots, k_i$) are the minimal centers of the Participants of sc_i.

For an output sentence s, $\mathbb{1}_s(u)$ is a function from the input units to $\{0, 1\}$, which returns 1 iff u is aligned (according to A) with a word in s.[4]

The role of the non-splitting penalty term n_{out}/n_{inp} in the SAMSA formula is to penalize cases where the number of sentences in the output is smaller than the number of Scenes. In order to isolate the effect of the non-splitting penalty, we experiment with an additional metric SAMSA_{abl} (reads "SAMSA ablated"), which is identical to SAMSA but does not take this term into account. Corpus-level SAMSA and SAMSA_{abl} scores are obtained by averaging their sentence scores.

In the case of implicit units i.e. omitted units that do not appear explicitly in the text (Abend and Rappoport, 2013), since the unit preservation cannot be directly captured, the score t for the relevant unit will be set to 0.5. For example, in the Scene "traveling is fun", the people who are traveling correspond to an implicit Participant. As implicit units are not covered by TUPA, this will only be relevant for the semi-automatic implementation of the metric (see Section 6).

5 Human Evaluation Benchmark

5.1 Evaluation Protocol

For testing the automatic metric, we first build a human evaluation benchmark, using 100 sentences from the complex part of the PWKP

[4] In some cases, the unit u can be a sequence of centers (if there are several minimal centers). In these cases, $\mathbb{1}_s(u)$ returns 1 iff the condition holds for all centers.

corpus and the outputs of six recent simplification systems for these sentences:[5] (1) TSM (Zhu et al., 2010) using Tree-Based SMT, (2) RevILP (Woodsend and Lapata, 2011) using Quasi-Synchronous Grammars, (3) PBMT-R (Wubben et al., 2012) using Phrase-Based SMT, (4) Hybrid (Narayan and Gardent, 2014), a supervised system using DRS, (5) UNSUP (Narayan and Gardent, 2016), an unsupervised system using DRS, and (6) Split-Deletion (Narayan and Gardent, 2016), the unsupervised system with only structural operations.

All these systems explicitly address at least one type of structural simplification operation. The last system, Split-Deletion, performs only structural (i.e., no lexical) operations. It is thus an interesting test case for SAMSA since here the aligner can be replaced by a simple match between identical words. In total we obtain 600 system outputs from the six systems, as well as 100 sentences from the simple Wikipedia side of the corpus, which serve as references. Five in-house annotators with high proficiency in English evaluated the resulting 700 input-output pairs by answering the questions in Table 1.[6]

Qa addresses grammaticality, Qb and Qc capture two complementary aspects of meaning preservation (the addition and the removal of information) and Qd addresses structural simplicity. Possible answers are: 1 ("no"), 2 ("maybe") and 3 ("yes"). Following Glavas and Štajner (2013), we used a 3 point Likert scale, which has recently been shown to be preferable over a 5 point scale through human studies on sentence compression (Toutanova et al., 2016).

Question Qd was accompanied by a negative example [7] showing a case of lexical simplification, where a complex word is replaced by a simple one. A positive example was not included so as not to bias the annotators by revealing the nature of the operations our experiments focus on (i.e., splitting and deletion).

The PWKP test corpus (Zhu et al., 2010) was selected for our experiments over the development and test sets used in (Xu et al., 2016), as the latter's selection process was explicitly biased towards input-output pairs that mainly contain lexical simplifications.

Qa	Is the output grammatical?
Qb	Does the output add information, compared to the input?
Qc	Does the output remove important information, compared to the input?
Qd	Is the output simpler than the input, ignoring the complexity of the words?

Table 1: Questions for the human evaluation

5.2 Human Score Computation

Given the annotator's answers, we consider the following scores. First, the grammaticality score $\mathcal{G}$ is the answer to Qa. By inverting (changing 1 to 3 and 3 to 1) the answer for Qb, we obtain a Non-Addition score indicating to which extent no additional information has been added. Similarly, inverting the answer to Qc yields the Non-Removal score. Averaging these two scores, we obtain the meaning preservation score $\mathcal{P}$. Finally, the structural simplicity score $\mathcal{S}$ is the answer to Qd. Each of these scores is averaged over the five annotators. We further compute an average human score:

$$\text{AvgHuman} = \frac{1}{3}(\mathcal{G} + \mathcal{P} + \mathcal{S})$$

5.3 Inter-annotator Agreement

Inter-annotator agreement rates are computed in two ways. Table 2 presents the absolute agreement and Cohen's quadratic weighted κ (Cohen, 1968). Table 3 presents Spearman's correlation (ρ) between the human ratings of the input-output pairs (top row), and between the resulting system scores (bottom row). In both cases, the agreement between the five annotators is computed as the average agreement over the 10 annotator pairs.

	Qa	Qb	Qc	Qd
Total	0.58 (0.56)	0.74 (0.30)	0.53 (0.45)	0.57 (0.10)
TSM	0.59 (0.47)	0.75 (0.27)	0.50 (0.40)	0.43 (0.08)
RevILP	0.61 (0.59)	0.78 (0.27)	0.60 (0.43)	0.62 (0.11)
PBMT-R	0.47 (0.42)	0.70 (0.20)	0.58 (0.31)	0.76 (0.10)
Hybrid	0.59 (0.46)	0.77 (0.26)	0.52 (0.48)	0.72 (0.15)
UNSUP	0.51 (0.42)	0.59 (0.10)	0.45 (0.17)	0.52 (0.04)
Split-Deletion	0.59 (0.48)	0.93 (0.02)	0.45 (0.29)	0.55 (0.04)
Reference	0.70 (0.40)	0.66 (0.46)	0.52 (0.58)	0.41 (0.12)

Table 2: Inter-annotator absolute agreement (and quadratic weighted κ), averaged over the 10 annotator pairs. Rows correspond to systems, columns to questions. The top "Total" row refers to the concatenation of the outputs of all 6 systems together with the reference sentences.

6 Experimental Setup

We further compute SAMSA for the 100 sentences of the PWKP test set and the corresponding system outputs. Experiments are conducted in

[5] All the data can be found here: http://homepages.inf.ed.ac.uk/snaraya2/data/simplification-2016.tgz.

[6] Each input-output pair was rated by all five annotators.

[7] Other questions appeared without any example.

	Qa	Qb	Qc	Qd	AvgHuman
Sen.	0.63^{*}	0.30^{*}	0.48^{*}	0.11^{**}	0.49^{*}
Sys.	0.92^{**}	0.54 (0.1)	0.64 (0.06)	0.14 (0.4)	0.64 (0.06)

Table 3: Spearman's correlation (and p-values) of the system-level (top row) and sentence-level (bottom row) ratings of the five annotators. $^{*}p < 10^{-5}$, $^{**}p = 0.002$.

two settings: (1) a semi-automatic setting where UCCA annotation was carried out manually by a single expert UCCA annotator using the UCCAApp annotation software (Abend et al., 2017), and according to the standard annotation guidelines;[8] (2) an automatic setting where the UCCA annotation was carried out by the TUPA parser (Hershcovich et al., 2017). Sentence segmentation of the outputs was carried out using the NLTK package (Loper and Bird, 2002). For word alignments, we used the aligner of Sultan et al. (2014).[9]

7 Correlation with Human Evaluation

We compare the system rankings obtained by SAMSA and by the four human parameters. We find that the two leading systems according to AvgHuman and SAMSA turn out to be the same: Split-Deletion and RevILP. This is the case both for the semi-automatic and the automatic implementations of the metric. A Spearman ρ correlation between the human and SAMSA scores (comparing their rankings) is presented in Table 4.

We compare SAMSA and SAMSA$_{abl}$ to the reference-based measures SARI[10] (Xu et al., 2016) and BLEU, as well as to the negative Levenshtein distance to the reference (-LD$_{\text{SR}}$). We use the only available reference for this corpus, in accordance with the standard practice. SARI is a reference-based measure, based on n-gram overlap between the source, output and reference, and focuses on lexical (rather than structural) simplification. For completeness, we include the other two measures reported in Narayan and Gardent (2016), which are measures of similarity to the input (i.e., they quantify the tendency of the systems to introduce changes to the input): the negative Levenshtein distances between the output and input compared to the original complex corpus (-LD$_{\text{SC}}$), and the number of sentences split by each of the systems.

[8] http://www.cs.huji.ac.il/~oabend/ucca.html

[9] https://github.com/ma-sultan/monolingual-word-aligner

[10] Data and code for can be found in https://github.com/cocoxu/simplification.

The highest correlation with AvgHuman and grammaticality is obtained by semi-automatic SAMSA (0.58 and 0.54), a high correlation especially in comparison to the inter-annotator agreement on AvgHuman (0.64, Table 3). The automatic version obtains high correlation with human judgments in these settings, where for structural simplicity, it scores somewhat higher than the semi-automatic SAMSA. The highest correlation with structural simplicity is obtained by the number of sentences with splitting, where SAMSA (automatic and semi-automatic) is second and third highest, although when restricted to multi-Scene sentences, the correlation for SAMSA (semi-automatic) is higher (0.89, $p = 0.009$ and 0.77, $p = 0.04$).

The highest correlation for meaning preservation is obtained by SAMSA$_{abl}$ which provides further evidence that the retainment of semantic structures is a strong predictor of meaning preservation (Sulem et al., 2015). SAMSA in itself does not correlate with meaning preservation, probably due to its penalization of under-splitting sentences.

Note that the standard reference-based measures for simplification, BLEU and SARI, obtain low and often negative correlation with human ratings. We believe that this is the case because SARI and BLEU admittedly focus on lexical simplification, and are difficult to use to rank systems which also perform structural simplification.

Our results thus suggest that SAMSA provides additional value in predicting the quality of a simplification system and should be reported in tandem with more lexically-oriented measures.

8 Discussion

Human evaluation parameters. The fact that the highest correlations for structural simplicity and meaning preservation are obtained by different metrics (SAMSA and SAMSA$_{abl}$ respectively) highlights the complementarity of these two parameters for evaluating TS quality but also the difficulty of capturing them together. Indeed, a given sentence-level operation could both change the original meaning by adding or removing information (affecting the $\mathcal{P}$ score) and increase simplicity ($\mathcal{S}$). On the other hand, the identity transformation perfectly preserves the meaning of the original sentence without making it simpler.

For examining this phenomenon, we compute Spearman's correlation at the system-level between the simplicity and meaning preservation hu-

	Reference-less				Reference-based			Δ from Source	
	SAMSA		SAMSA$_{abl}$		BLEU	SARI	-LD$_{SR}$	-LD$_{SC}$	# Split Sents.
	Semi-Auto.	Auto.	Semi-Auto.	Auto.					
$\mathcal{G}$	**0.54** (0.1)	0.37 (0.2)	0.14 (0.4)	0.14 (0.4)	0.09 (0.4)	-0.77 (0.04)	-0.43 (0.2)	-0.09 (0.4)	0.09 (0.4)
$\mathcal{P}$	-0.09 (0.4)	-0.37 (0.2)	**0.54** (0.1)	**0.54** (0.1)	0.37 (0.2)	-0.14 (0.4)	0.03 (0.5)	0.37 (0.2)	-0.49 (0.2)
$\mathcal{S}$	0.54 (0.1)	0.71 (0.06)	-0.71 (0.06)	-0.71 (0.06)	-0.60 (0.1)	-0.43 (0.2)	-0.43 (0.2)	-0.54 (0.1)	**0.83** (0.02)
AvgHuman	**0.58** (0.1)	0.35 (0.1)	0.09 (0.2)	0.09 (0.2)	0.06 (0.5)	-0.81 (0.02)	-0.46 (0.2)	-0.12 (0.4)	0.14 (0.4)

Table 4: Spearman's correlation of system scores i.e. Pearson's correlation of system rankings (and p-values), between evaluation measures (columns) and human judgments (rows). The ranking is between the six simplification systems experimented with. The left block of columns corresponds to the SAMSA and SAMSA$_{abl}$ measures, in their semi-automatic and automatic forms. The middle block of columns corresponds to the reference-based measures SARI and BLEU, as well as -LD$_{\mathrm{SR}}$, which is the negative Levenshtein distances of the system output from the reference. The right block corresponds to measures of conservatism, and reflect how well the tendency of the systems to introduce changes to the input correlates with the human rankings. The block includes -LD$_{\mathrm{SC}}$, the negative Levenshtein distance from the source sentence, and the number of input sentences split by each of the systems. Levenshtein distances are taken as negative in order to capture similarity between the output and source/reference. The measure with the highest correlation in each row is boldfaced.

man scores. We obtain a correlation of -0.77 ($p = 0.04$) between $\mathcal{S}$ and $\mathcal{P}$. The correlation between $\mathcal{S}$ and the two sub-components of $\mathcal{P}$, the Non-Addition and the Non-Removal scores, are -0.43 ($p = 0.2$) and -0.77 ($p = 0.04$) respectively. These negative correlations support our use of an average human score for assessing the overall quality of the simplification.

Distribution at the sentence level. In addition to the system-level analysis presented in Section 7, we also investigate the behavior of SAMSA at the sentence level by examining its joint distribution with the human evaluation scores. Focusing on the AvgHuman score and the automatic implementation of SAMSA and using the same data as in Section 7, we consider a single pair of scores $(\mathrm{AvgHuman}_i, \mathrm{SAMSA}_i)$, $1 \leq i \leq 100$, for each of the 100 source sentences, averaging over the SAMSA and human scores obtained for the 6 simplification systems (See Figure 1).

The joint distribution indicates a positive correlation between SAMSA and AvgHuman. The corresponding Pearson correlation is indeed 0.27 ($p = 0.03$).

9 Evaluation on the QATS Benchmark

In order to provide further validation for SAMSA predictive value for quality of simplification systems, we report SAMSA's correlation with a recently proposed benchmark, used for the QATS (Quality Assessment for Text Simplification) shared task (Štajner et al., 2016).

Setup. The test corpus contains 126 sentences taken from 3 datasets described in Štajner et al. (2016)[11]: (1) EventS: original sentences from the EMM News-Brief[12] and their syntactically simplified versions (with significant content reduction) by the EventSimplify TS system (Glavas and Štajner, 2013)[13] (the test corpus contains 54 pairs from this dataset), (2) EncBrit: original sentences from the Encyclopedia Britannica (Barzilay and Elhadad, 2003) and their automatic simplifications obtained using ATS systems based on several phrase-based statistical MT systems (Štajner et al., 2015) trained on Wikipedia TS corpus (Coster and Kauchak, 2011) (24 pairs), and (3) LSLight: sentences from English Wikipedia and their automatic simplifications (Glavaš and Štajner, 2015) by three different lexical simplification systems (Biran et al., 2011; Horn et al., 2014; Glavaš and Štajner, 2015) (48 pairs).

Human evaluation is also provided by this resource, with scores for overall quality, grammaticality, meaning preservation and simplicity. Importantly, the simplicity score does not explicitly refer to the output's structural simplicity, but rather to its readability. We focus on the overall human score, and compare it to SAMSA. Since different systems were used to simplify different portions of the input, correlation is taken at the sentence level.

We use the same implementations of SAMSA. Manual UCCA annotation is here performed by one of the authors of this paper.

Results. We follow Štajner et al. (2016) and report the Pearson correlations (at the sentence

[11] http://qats2016.github.io/shared.html

[12] emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html

[13] takelab.fer.hr/data/symplify

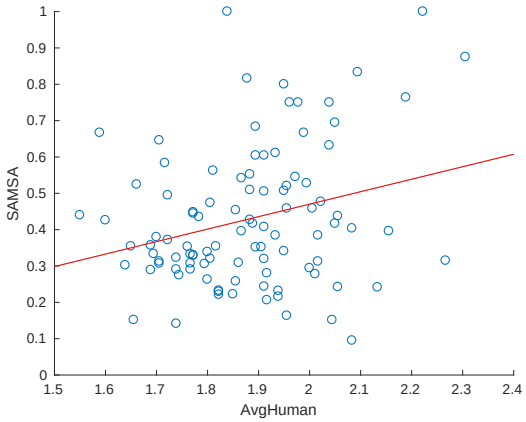

Figure 1: Joint distribution of the automatic SAMSA and the AvgHuman scores at the sentence level. Each point in the graph corresponds to a single source sentence. In addition to the scatter plot, a least-squares regression line is presented.

level) between the rankings of the metrics and the human evaluation scores. Results show that the semi-automatic/automatic SAMSA obtains a Pearson correlation of 0.32 and 0.28 with the human scores. This places these measures in the 3rd and 4th places in the shared task, where the only two systems that surpassed it are marginally better, with scores of 0.33 and 0.34, and where the next system in QATS obtained a correlation of 0.23.

This correlation by SAMSA was obtained in more restricted conditions, compared to the measures that competed in QATS. First, SAMSA computes its score by only considering the UCCA structure of the source, and an automatic word-to-word alignment between the source and output. Most QATS systems, including OSVCML and OSVCML2 (Nisioi and Nauze, 2016) which scored highest on the shared task, use an ensemble of classifiers based on bag-of-words, POS tags, sentiment information, negation, readability measures and other resources. Second, the systems participating in the shared task had training data available to them, annotated by the same annotators as the test data. This was used to train classifiers for predicting their score. This gives the QATS measures much predictive strength, but hampers their interpretability. SAMSA on the other hand is conceptually simple and interpretable. Third, the QATS shared task does not focus on structural simplification, but experiments on different types of systems. Indeed, some of the data was annotated by systems that exclusively perform lexical simplification, which is orthogonal to SAMSA's structural focus.

Given these factors, SAMSA's competitive correlation with the participating systems in QATS suggests that structural simplicity, as reflected by the correct splitting of UCCA Scenes, captures a major component in overall simplification quality, underscoring SAMSA's value. These promising results also motivate a future combination of SAMSA with classifier-based metrics.

10 Conclusion

We presented the first structure-aware metric for text simplification, SAMSA, and the first evaluation experiments that directly target the structural simplification component, separately from the lexical component. We argue that the structural and lexical dimensions of simplification are loosely related, and that TS evaluation protocols should assess both. We empirically demonstrate that strong measures that assess lexical simplification quality (notably SARI), fail to correlate with human judgments when structural simplification is performed by the evaluated systems. Our experiments show that SAMSA correlates well with human judgments in such settings, which demonstrates its usefulness for evaluating and tuning statistical simplification systems, and shows that structural evaluation provides a complementary perspective on simplification quality.

Acknowledgments

We would like to thank Zhemin Zhu and Sander Wubben for sharing their data, as well as the annotators for participating in our evaluation and UCCA annotation experiments. We also thank Daniel Hershcovich and the anonymous review-

ers for their helpful comments. This work was partially supported by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI) and by the Israel Science Foundation (grant No. 929/17), as well as by the HUJI Cyber Security Research Center in conjunction with the Israel National Cyber Bureau in the Prime Minister's Office.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proc. of ACL-13*. pages 228–238. http://aclweb.org/anthology/P/P13/P13-1023.pdf.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proc. of ACL'17*. pages 77–89. http://aclweb.org/anthology/P/P17/P17-1008.pdf.

Omri Abend, Shai Yerushalmi, and Ari Rappoport. 2017. UCCAApp: Web-application for syntactic and semantic phrase-based annotation. In *Proc. of ACL'17, System Demonstrations*. pages 109–114. http://www.aclweb.org/anthology/P17-4019.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Rrepresentation for sembanking. *Proc. of Linguistic Annotation Workshop and Interoperability with Discourse* pages 178–186. http://aclweb.org/anthology/W/W13/W13-2322.pdf.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proc. of EMNLP'03*. pages 25–32. http://www.aclweb.org/anthology/W/W03/W03-1004.pdf.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proc. of ACL'11*. pages 465–501. http://aclweb.org/anthology/P/P11/P11-2087.pdf.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-based evaluation of machine translation. In *Proc. of EMNLP'16*. pages 1264–1274. http://aclweb.org/anthology/D/D16/D16-1134.pdf.

Yvonne Margaret Canning. 2002. *Syntactic simplification of text*. Ph.D. thesis, University of Sunderland, UK.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for sentence simplification. In *Proc. of COLING'96*. pages 1041–1044. http://aclweb.org/anthology/C/C96/C96-2183.pdf.

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. In *Proc. of NAACL'18 (Short papers)*. To appear.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proc. of ACL-COLING'06*. pages 377–384. http://aclweb.org/anthology/P/P06/P06-1048.pdf.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proc. of ACL'11*. pages 665–669. http://aclweb.org/anthology/P/P11/P11-2117.pdf.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proc. of WMT'11*. pages 85–91. http://aclweb.org/anthology/W/W11/W11-2107.pdf.

Robert M.W. Dixon. 2010a. *Basic Linguistic Theory: Grammatical Topics*, volume 2. Oxford University Press.

Robert M.W. Dixon. 2010b. *Basic Linguistic Theory: Methodology*, volume 1. Oxford University Press.

Robert M.W. Dixon. 2012. *Basic Linguistic Theory: Further Grammatical Topics*, volume 3. Oxford University Press.

Goran Glavas and Sanja Štajner. 2013. Event-centered simplification of news stories. In *Proc. of the Student Research Workshop associated with RANLP 2013*. pages 71–78. http://aclweb.org/anthology/R13-2011.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proc. of ACL'15 (Short papers)*. pages 63–68. http://www.aclweb.org/anthology/P15-2011.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In *Proc. of ACL'17*. pages 1127–1138. http://aclweb.org/anthology/P/P17/P17-1104.pdf.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proc. of ACL'14 (Short papers)*. pages 458–463. http://aclweb.org/anthology/P/P14/P14-2075.pdf.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from Standard Wikipedia to Simple Wikipedia. In *Proc. of NAACL'15*. pages 211–217. http://aclweb.org/anthology/N/N15/N15-1022.pdf.

Hans Kamp. 1981. A theory of truth and semantic representation. In J.A.G. Groenendijk, T.M.V. Jassen, B.J. Stokhof, and M.J.B/ Stokhof, editors, *Formal methods in the study of language*. Mathematisch Centrum. Number pt.1 in Mathematical Centre tracts.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical report, DTIC Document.

Ronald W. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, USA.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic mt evaluation without reference translations. In *Proc. of ACL'14 (Short Papers)*. pages 765–771. http://aclweb.org/anthology/P/P14/P14-2124.pdf.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic MT evaluation. In *Proc. of WMT'12*. pages 243–252. http://aclweb.org/anthology/W/W12/W12-3129.pdf.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proc. of EMNLP'02*. pages 63–70. http://www.aclweb.org/anthology/W/W02/W02-0109.pdf.

Kshitij Mishra, Ankush Soni, Rahul Sharma, and Dipti Misra Sharma. 2014. Exploring the effects of sentence simplification on Hindi to English Machine Translation systems. In *Proc. of the Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society*. pages 21–29. http://www.aclweb.org/anthology/W14-5603.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proc. of ACL14*. pages 435–445. http://aclweb.org/anthology/P/P14/P14-1041.pdf.

Shashi Narayan and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proc. of INLG'16*. pages 111–120. http://aclweb.org/anthology/W/W16/W16-6620.pdf.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proc. of EMNLP'17*. pages 617–627. http://aclweb.org/anthology/D/D17/D17-1065.pdf.

Christina Niklaus, Bernahard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proc. of COLING'16*. http://aclweb.org/anthology/C/C16/C16-2036.pdf.

Sergiu Nisioi and Fabrice Nauze. 2016. An ensemble method for quality assessment of text simplification. In *Workshop on Quality Assessment for Text Simplification (QATS)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL'02*. pages 311–318. http://aclweb.org/anthology/P/P02/P02-1040.pdf.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help?: text simplification strategies for people with dyslexia. In *Proc. of the 10th International Cross-Disciplinary Conference on Web Accesibility*. pages 15:1 – 15:10.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. TINE: A metric to assess MT adequacy. In *Proc. of WMT'11*. pages 116–122. http://aclweb.org/anthology/W/W11/W11-2112.pdf.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Proc. of LEC*. pages 64–71.

Advaith Siddharthan and M. A. Angrosh. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proc. of EACL'14*. pages 722–731. http://aclweb.org/anthology/E/E14/E14-1076.pdf.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proc. of WMT'09*. pages 85–91. http://www.aclweb.org/anthology/W09-0441.

Elior Sulem, Omri Abend, and Ari Rappoport. 2015. Conceptual annotations preserve structure across translations. In *Proc. of 1st Workshop on Semantics-Driven Statistical Machine Translation (S2Mt 2015)*. pages 11–22. http://aclweb.org/anthology/W/W15/W15-3502.pdf.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *TACL* 2:219–230. http://aclweb.org/anthology/Q/Q14/Q14-1018.pdf.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proc. of ACL'12*. pages 38–42. http://aclweb.org/anthology/P/P12/P12-2008.pdf.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proc. of EMNLP'16*. pages 340–350. http://aclweb.org/anthology/D/D16/D16-1033.pdf.

Sanja Štajner, Hannah Bechara, and Horacio Saggion. 2015. A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In *Proc. of ACL'15, Short papers*. pages 823–828. http://aclweb.org/anthology/P/P15/P15-2135.pdf.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. *Proc. of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations* pages 1–10. http://www.aclweb.org/anthology/W14-1201.

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification. In *Workshop on Quality Assessment for Text Simplification (QATS)*.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proc. of EMNLP'11*. pages 409–420. http://aclweb.org/anthology/D/D11/D11-1038.pdf.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proc. of ACL'12*. pages 1015–1024. http://aclweb.org/anthology/P/P12/P12-1107.pdf.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: new data can help. *TACL* 3:283–297. http://aclweb.org/anthology/Q/Q15/Q15-1021.pdf.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *TACL* 4:401–415. http://aclweb.org/anthology/Q/Q16/Q16-1029.pdf.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proc. of COLING'10*. pages 1353–1361. http://aclweb.org/anthology/C/C10/C10-1152.pdf.