Semantic Structural Evaluation for Text Simplification

Elior Sulem, Omri Abend and Ari Rappoport

The Hebrew University of Jerusalem

NAACL HLT 2018
Text Simplification

Last year I read the book John authored → John wrote a book. I read the book.

Original sentence One or several simpler sentences
Text Simplification

Last year I read the book John authored

Original sentence

John wrote a book. I read the book.

One or several simpler sentences

Multiple motivations

Preprocessing for Natural Language Processing tasks

e.g., machine translation, relation extraction, parsing

Reading aids, Language Comprehension

e.g., people with aphasia, dyslexia, second language learners
Two types of Simplification

Last year I read the book John authored → John wrote a book. I read the book.

Original sentence → One or several simpler sentences

Lexical operations

e.g., word substitution

Structural operations

e.g., sentence splitting, deletion

All the previous evaluation approaches targeted lexical simplification.

Here: the first automatic evaluation measure for structural simplification.
Overview

1. Current Text Simplification Evaluation
2. A New Measure for Structural Simplification
   - **SAMSA** (Simplification Automatic Measure through Semantic Annotation)
     - 2.1. SAMSA properties
     - 2.2. The semantic structures
     - 2.3. SAMSA computation
3. Human Evaluation Benchmark
4. Correlation Analysis with Human Evaluation
5. Conclusion
Current Text Simplification Evaluation

Main automatic metrics

BLEU, Panini et al., 2002
SARI, Xu et al., 2016

Reference-based
The output is compared to one or multiple references

Focus on lexical aspects
Do not take into account structural aspects
A New Measure for Structural Simplification

SAMSA

Simplification Automatic evaluation Measure through Semantic Annotation
SAMSA Properties

- Measures the **preservation of the sentence-level semantics**
- Measures **structural simplicity**
- No reference simplifications
- Fully automatic
- Semantic parsing only on the source side
SAMSA Properties

Example:

John arrived home and gave Mary a call. (input)

John arrived home. John called Mary. (output)

Assumption:

In an ideal simplification each event is placed in a different sentence.

Fits with existing practices in Text Simplification.

(Glavaš and Štajner, 2013; Narayan and Gardent, 2014)
SAMSA Properties

Example:

John arrived home and gave Mary a call. (input)

John arrived home. John called Mary. (output)

**SAMSA** focuses on the core semantic components of the sentence, and is tolerant to the deletion of other units.
The Semantic Structures

Semantic Annotation: **UCCA** (Abend and Rappoport, 2013)

- Based on typological and cognitive theories

(Dixon, 2010, 2012; Langacker, 2008)

<table>
<thead>
<tr>
<th>Process (P)</th>
<th>Function (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant (A)</td>
<td>Parallel Scene (H)</td>
</tr>
<tr>
<td>Center (C)</td>
<td>Linker (L)</td>
</tr>
<tr>
<td>Elaborator (E)</td>
<td>Relator (R)</td>
</tr>
</tbody>
</table>

John arrived home and gave a call to Mary.
The Semantic Structures

Semantic Annotation: **UCCA** (Abend and Rappoport, 2013)

- Stable across translations (Sulem, Abend and Rappoport, 2015)
- Used for the evaluation of MT and GEC (Birch et al., 2016; Choshen and Abend, 2018)

![Diagram of semantic annotation with labels Process (P), Function (F), Participant (A), Parallel Scene (H), Center (C), Linker (L), Elaborator (E), Relator (R)]

John arrived home and gave a call to Mary.
The Semantic Structures

Semantic Annotation: **UCCA** (Abend and Rappoport, 2013)

- Explicitly annotates semantic distinctions, abstracting away from syntax (like AMR; Banarescu et al., 2013)
- Unlike AMR, semantic units are directly anchored in the text.
The Semantic Structures

Semantic Annotation: **UCCA** (Abend and Rappoport, 2013)
- UCCA parsing (Hershcovich et al., 2017, 2018)
- Shared Task in Sem-Eval 2019!

```
John arrived home
and gave a call to Mary
```

<table>
<thead>
<tr>
<th>Process (P)</th>
<th>Function (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant (A)</td>
<td>Parallel Scene (H)</td>
</tr>
<tr>
<td>Center (C)</td>
<td>Linker (L)</td>
</tr>
<tr>
<td>Elaborator (E)</td>
<td>Relator (R)</td>
</tr>
</tbody>
</table>
The Semantic Structures

Semantic Annotation: UCCA (Abend and Rappoport, 2013)
- Scenes evoked by a Main Relation (Process or State).

```
John arrived home
and
Gave a call to Mary
```

Process (P) Function (F)
Participant (A) Parallel Scene (H)
Center (C) Linker (L)
Elaborator (E) Relator (R)
The Semantic Structures

Semantic Annotation: **UCCA** (Abend and Rappoport, 2013)

- A Scene may contain one or several **Participants**.

---

```
Process (P)  Function (F)
Participant (A)  Parallel Scene (H)
Center (C)  Linker (L)
Elaborator (E)  Relator (R)
```

---

**Diagram:**

- John arrived home
- and
- gave a call to Mary
SAMSA Computation

Example:

John arrived home  John gave Mary a call  (input Scenes)

John arrived home. John called Mary. (output sentences)

1. Match each Scene to a sentence.

2. Give a score to each Scene assessing its meaning preservation in the aligned sentence.
   ➔ Evaluated through the preservation of its main semantic components.

3. Average the scores and penalize non-splitting.
Scene to Sentence Matching:

- A word alignment tool is used (Sultan et al., 2014) for aligning a Scene to the candidate sentences.
  - Each word is aligned to 1 or 0 words in the candidate sentence.

- To each Scene we match the sentence for which the highest number of word alignments is obtained.

- If there are more sentences than Scenes, a score of zero is assigned.

John arrived home  John gave Mary a call  (input Scenes)

John arrived home.  John called Mary.  (output sentences)
**SAMSA Computation**

Word alignment

Scene: John gave Mary a call
Sentence: John called Mary

UCCA annotation

Suppose the Scene $Sc$ is matched to the sentence $Sen$:

$$\text{Score}_{Sen}(Sc) = \frac{1}{2} \left( \text{Score}_{Sen}(MR) + \frac{1}{K} \sum_{i=1}^{K} \text{Score}_{Sen}(Par_k) \right)$$

$MR$ - Minimal center of the Main Relation (Process / State)

$Par_k$ - Minimal center of the $k$th Participant

$$\text{Score}_{Sen}(u) = \begin{cases} 
1 & \text{if } u \text{ is aligned to a word in } Sen \\
0 & \text{otherwise}
\end{cases}$$
SAMSA Computation

• Average over the input Scenes

• Non-splitting penalty: \[ \frac{n_{out}}{n_{inp}} \]  
  - Number of output sentences
  - Number of input Scenes

\[ \text{We also experiment with SAMSA}_{abl}, \text{ without non-splitting penalty.} \]
Human Evaluation Benchmark

- 5 annotators

- 100 source sentences (PWKP test set)

- 6 Simplification systems + Simple corpus

- 4 Questions for each input-output pair (1 to 3 scale):

<table>
<thead>
<tr>
<th>Qa</th>
<th>Is the output grammatical?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qb</td>
<td>Does the output add information, compared to the input?</td>
</tr>
<tr>
<td>Qc</td>
<td>Does the output remove important information, compared to the input?</td>
</tr>
<tr>
<td>Qd</td>
<td>Is the output simpler than the input, ignoring the complexity of the words?</td>
</tr>
</tbody>
</table>

- Parameters: - Grammaticality (G)
  - Meaning Preservation (P)
  - Structural Simplicity (S)
Human Evaluation Benchmark

- 5 annotators

- 100 source sentences (PWKP test set)

- 6 Simplification systems + Simple corpus

- 4 Questions for each input-output pair (1 to 3 scale):

<table>
<thead>
<tr>
<th>Qa</th>
<th>Is the output grammatical?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qb</td>
<td>Does the output add information, compared to the input?</td>
</tr>
<tr>
<td>Qc</td>
<td>Does the output remove important information, compared to the input?</td>
</tr>
<tr>
<td>Qd</td>
<td>Is the output simpler than the input, ignoring the complexity of the words?</td>
</tr>
</tbody>
</table>

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

Human scores available at: https://github.com/eliorsulem/SAMSA
Correlation with Human Evaluation

Spearman’s correlation at the system level of the metric scores with the human evaluation scores, considering the output of the 6 simplification systems

- **G** – Grammaticality, **P** – Meaning Preservation, **S** – Structural Simplicity

<table>
<thead>
<tr>
<th></th>
<th>Reference-less</th>
<th>Reference-based</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAMSA Semi-Aut.</td>
<td>SAMSA Aut.</td>
<td>SAMSA\text{abl} Semi-Aut.</td>
<td>SAMSA\text{abl} Aut.</td>
<td>BLEU</td>
<td>SARI</td>
<td>Sent. with Splits</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>0.54</td>
<td>0.37</td>
<td>0.14</td>
<td>0.14</td>
<td>0.09</td>
<td>-0.77</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>-0.09</td>
<td>-0.37</td>
<td>0.54</td>
<td>0.54</td>
<td>0.37</td>
<td>-0.14</td>
<td>-0.49</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>0.54</td>
<td>0.71</td>
<td>-0.71</td>
<td>-0.71</td>
<td>-0.60</td>
<td>-0.43</td>
<td>0.83</td>
</tr>
<tr>
<td>AvgHuman</td>
<td><strong>0.58</strong></td>
<td>0.35</td>
<td>0.09</td>
<td>0.09</td>
<td>0.06</td>
<td>-0.81</td>
<td>0.14</td>
</tr>
</tbody>
</table>

- **SAMSA** obtained the best correlation for AvgHuman.
- **SAMSA\text{abl}** obtained the best correlation for Meaning Preservation.
Correlation with Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Reference-less</th>
<th>Reference-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAMSA Semi-Aut.</td>
<td>SAMSA Aut.</td>
</tr>
<tr>
<td></td>
<td>SAMSA abl Semi-Aut.</td>
<td>SAMSA abl Aut.</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>SARI</td>
</tr>
<tr>
<td>G</td>
<td>0.54</td>
<td>0.37</td>
</tr>
<tr>
<td>P</td>
<td>-0.09</td>
<td>-0.37</td>
</tr>
<tr>
<td>S</td>
<td>0.54</td>
<td>0.71</td>
</tr>
<tr>
<td>AvgHuman</td>
<td>0.58</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Spearman’s correlation at the system level of the metric scores with the human evaluation scores, considering the output of the 6 simplification systems

G – Grammaticality, P – Meaning Preservation, S – Structural Simplicity

- **SAMSA** is ranked second and third for Simplicity.

- When restricted to multi-Scene sentences, SAMSA Semi-Aut. has a correlation of **0.89** (p=0.009). For Sent. with Splits, it is 0.77 (p=0.04).
Correlation with Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Reference-less</th>
<th>Reference-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMSA Semi-Aut.</td>
<td>0.54</td>
<td>0.37</td>
</tr>
<tr>
<td>SAMSA Aut.</td>
<td>-0.09</td>
<td>-0.37</td>
</tr>
<tr>
<td>SAMSA&lt;sub&gt;abl&lt;/sub&gt; Semi-Aut.</td>
<td>0.54</td>
<td>0.71</td>
</tr>
<tr>
<td>SAMSA&lt;sub&gt;abl&lt;/sub&gt; Aut.</td>
<td>0.09</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Spearman’s correlation at the system level of the metric scores with the human evaluation scores, considering the output of the 6 simplification systems

G – Grammaticality, P – Meaning Preservation, S – Structural Simplicity

⇒ High similarity between the Semi-Automatic and the Automatic implementations. For SAMSA<sub>abl</sub>, the ranking is the same.
Correlation with Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Reference-less</th>
<th>Reference-based</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAMSA Semi-Aut.</td>
<td>SAMSA Aut.</td>
<td>SAMSA_{abl} Semi-Aut.</td>
</tr>
<tr>
<td>G</td>
<td>0.54</td>
<td>0.37</td>
<td>0.14</td>
</tr>
<tr>
<td>P</td>
<td>-0.09</td>
<td>-0.37</td>
<td>0.54</td>
</tr>
<tr>
<td>S</td>
<td>0.54</td>
<td>0.71</td>
<td>-0.71</td>
</tr>
<tr>
<td>AvgHuman</td>
<td>0.58</td>
<td>0.35</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Spearman’s correlation **at the system level** of the metric scores with the human evaluation scores, considering the output of the **6 simplification systems**

- **G** – Grammaticality, **P** – Meaning Preservation, **S** – Structural Simplicity

⇒ Low and negative correlations for BLEU and SARI.
Correlation with Existing Benchmark

QATS task (Štajner et al., 2016)

Pearson Correlation with the Overall Human Score:

- Semi-automatic and automatic SAMSA rank 3\textsuperscript{rd} and 4\textsuperscript{th} (0.32 and 0.28), out of 15 measures.

- Surpassed by the best performing systems by a small margin (0.33 and 0.34).

Although: - We did not use training data (human scores)

- SAMSA focuses on structural simplicity.
Conclusion

- We proposed SAMSA, the **first structure-aware measure** for Text Simplification.

- SAMSA explicitly targets the **structural component** of Text Simplification.

- SAMSA gets **substantial correlations** with human evaluation.

- Existing measures fail to correlate with human judgments when structural simplification is performed.
Future Work

- SAMSA can be used for **tuning** Text Simplification systems.

- **Semantic decomposition** with UCCA can be used for improving Text Simplification (Sulem, Abend and Rappoport, ACL 2018).

- SAMSA can be extended to **other Text-to-Text generation tasks** as paraphrasing, sentence compression, or fusion.
Thank you

Elior Sulem

**Code and Data:**  https://github.com/eliorsulem/SAMSA

eliors@cs.huji.ac.il

www.cs.huji.ac.il/~eliors