Semantic Structural Evaluation for Text Simplification

Elior Sulem, Omri Abend and Ari Rappoport

The Hebrew University of Jerusalem

4th Usage-Based Linguistics Conference - July 2018

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Semantic Structural Evaluation for Text Simplification



Elior Sulem | Omri Abend | Ari Rappoport

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Based on a paper published in NAACL 2018.

Semantic Structural Evaluation for Text Simplification

Elior Sulem, Omri Abend and Ari Rappoport

Proc. of NAACL 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

Reduces the complexity of the sentence while **retaining its meaning.**

Text Simplification

Simplified Texts and Comprehension

Evidence that **simplified texts** can improve **reading comprehension**.

e.g., Mason & Kendall (1979); L'Allier (1980); Beck et al. (1991); Anderson & Davison (1988)

- Word substitutions

- Sentence splitting

- Making discourse relations explicit

- Information in cause-effect order

Text Simplification

Various Target Populations:

- **Non-native speakers** (e.g., Siddharthan et al., 2002)
- **Deafness** (e.g., Robbins and Hatcher, 1981)
- **Aphasia** (e.g., Caroll et al., 1999)
- **Dyslexia** (e.g., Rello et al., 2013)

Text Simplification

Natural Language Processing Applications:

- **Machine translation** (e.g., Mishra et al., 2014)
- **Parsing** (e.g., Chandrasekar et al., 1996)
- **Relation extraction** (e.g., Niklaus et al., 2016)

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

Automatic text simplification (rule-based or corpus-based):

word substitution, sentence splitting and deletion.

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, Panineni et al., 2002

SARI, Xu et al., 2016

➡ Reference-based

The output is compared to one or multiple references

➡ Focus on lexical aspects

Based on n-gram overlapping

Current Text Simplification Evaluation

N-gram overlapping

Candidate: Mary went to school.

Reference: John went home.

1-gram matches: 1

2-gram matches: 0

Current Text Simplification Evaluation

N-gram overlapping

Candidate: John went to school.

Reference: John went home.

1-gram matches: 2

2-gram matches: 1

Current Text Simplification Evaluation

Main automatic metrics

BLEU, Panineni et al., 2002

SARI, Xu et al., 2016

➡ Reference-based

Many possible simplifications for a given sentence

➡ 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)

score

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)

score

John arrived home. John called Mary. (output)

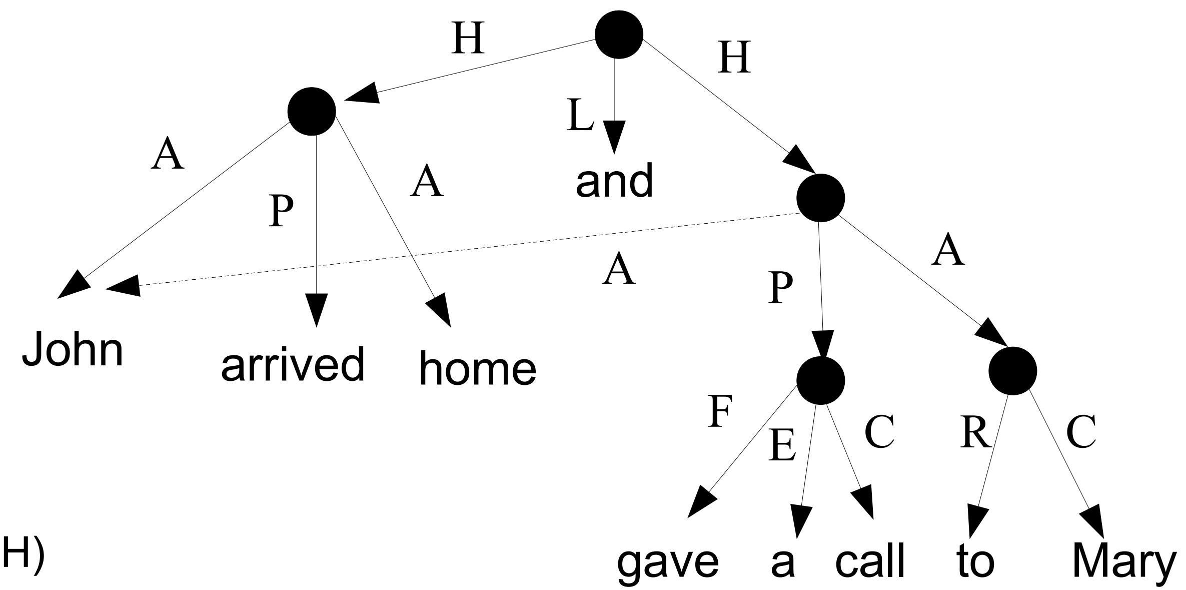
SAMSA focuses on the core semantic components of the sentence, and is tolerant to the deletion of other units.

The Semantic Structures

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

- Based on typological and cognitive theories

(Dixon, 2010, 2012; Langacker, 2008)

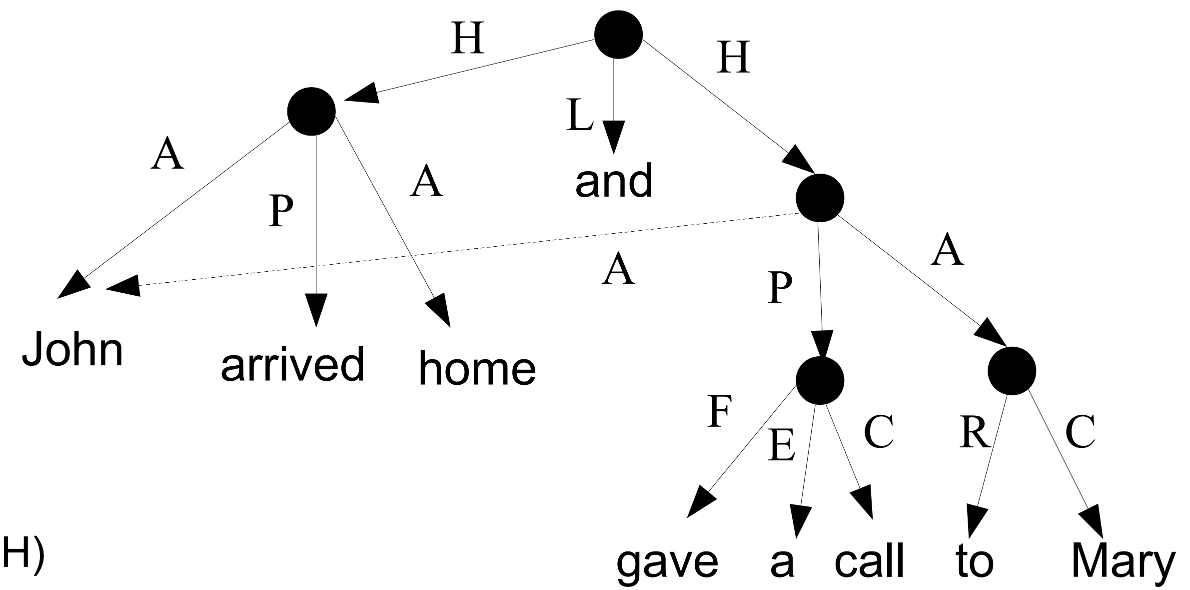


Process (P)	Function (F)
Participant (A)	Parallel Scene (H)
Center (C)	Linker (L)
Elaborator (E)	Relator (R)

The Semantic Structures

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

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



Process (P) Function (F)

Participant (A) Parallel Scene (H)

Center (C) Linker (L)

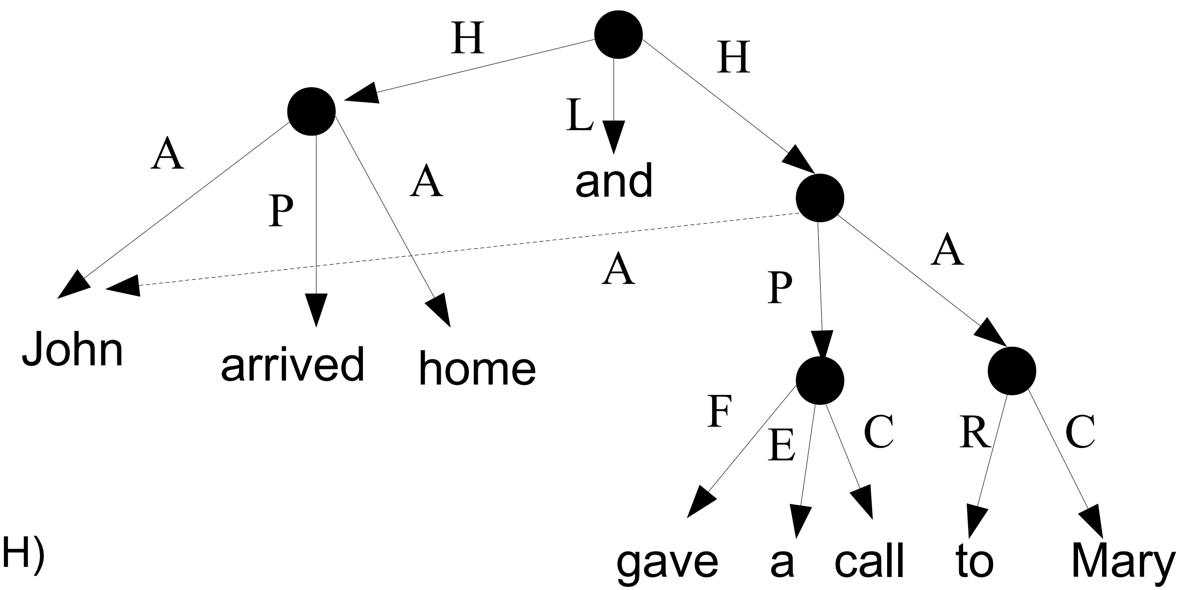
Elaborator (E) Relator (R)

The Semantic Structures

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

- Explicitly annotates semantic distinctions, abstracting away from syntax (like AMR; Banarescu et al., 2013)

- Unlike AMR, semantic units are directly anchored in the text.



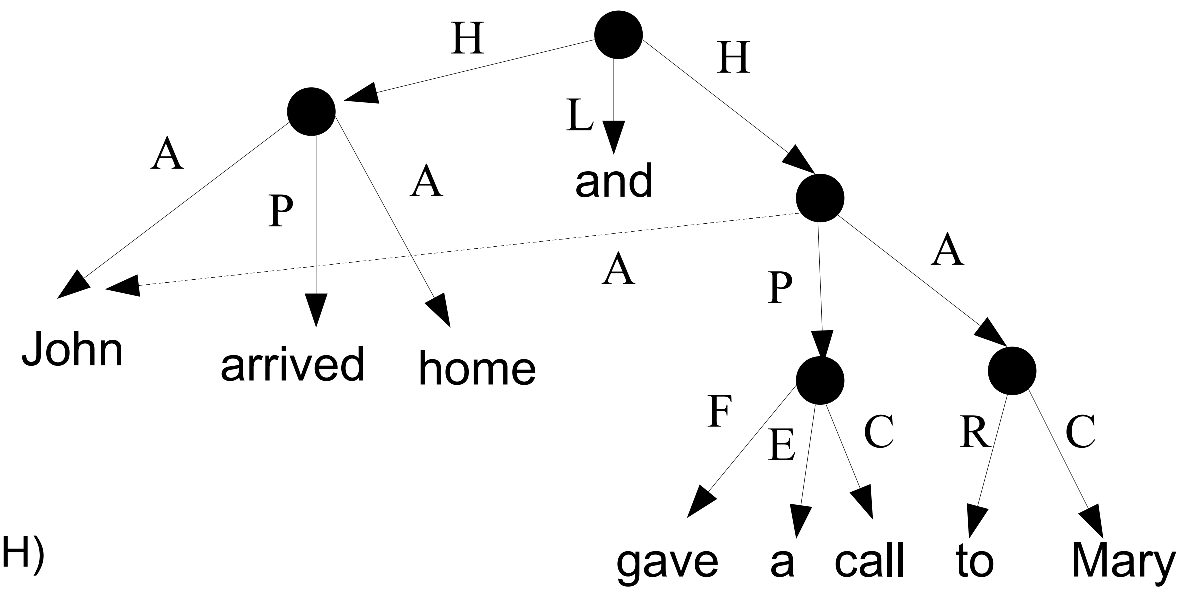
Process (P)	Function (F)
Participant (A)	Parallel Scene (H)
Center (C)	Linker (L)
Elaborator (E)	Relator (R)

The Semantic Structures

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

- UCCA parsing (Hershcovich et al., 2017, 2018)

TUPA parser – learning on a manually-annotated corpus



Process (P) Function (F)

Participant (A) Parallel Scene (H)

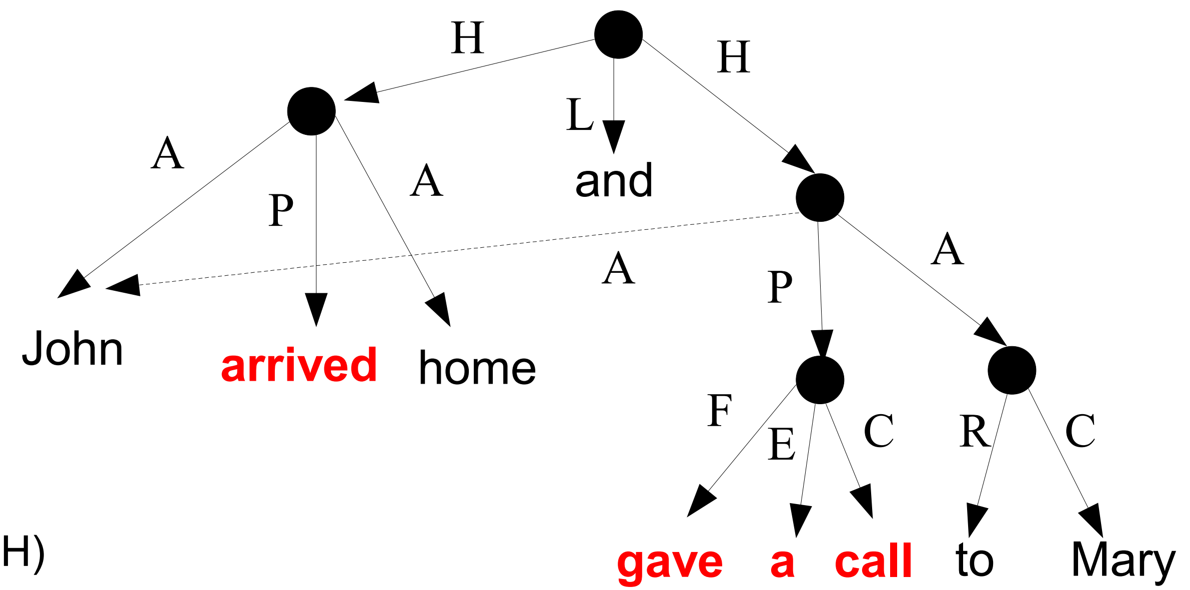
Center (C) Linker (L)

Elaborator (E) Relator (R)

The Semantic Structures

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

- **Scenes** evoked by a **Main Relation** (Process or State).

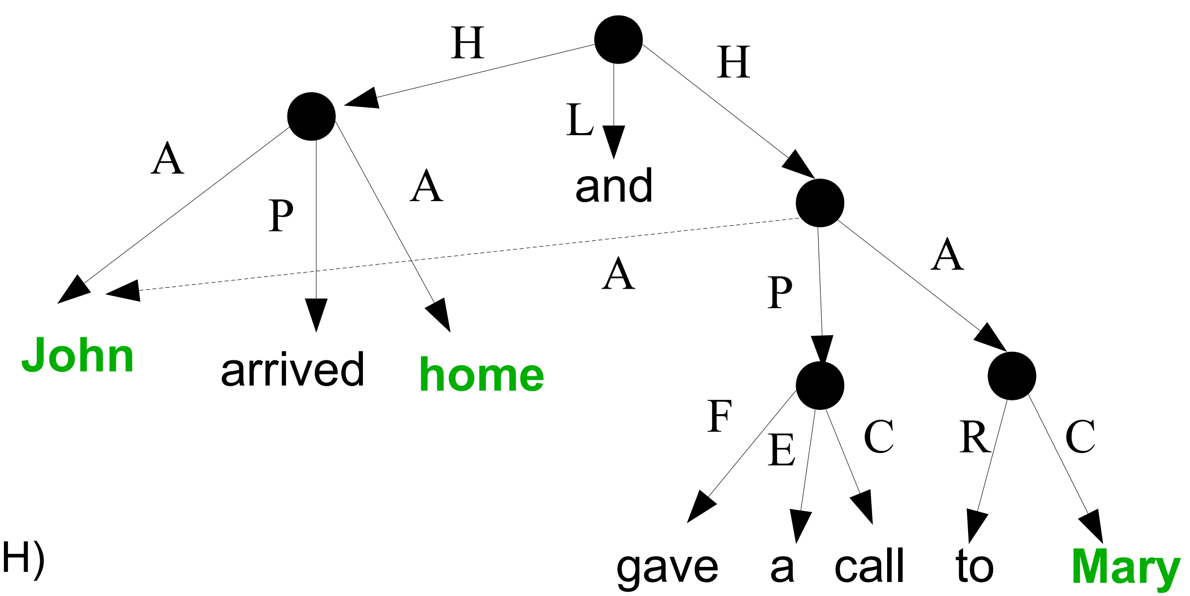


Process (P)	Function (F)
Participant (A)	Parallel Scene (H)
Center (C)	Linker (L)
Elaborator (E)	Relator (R)

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)

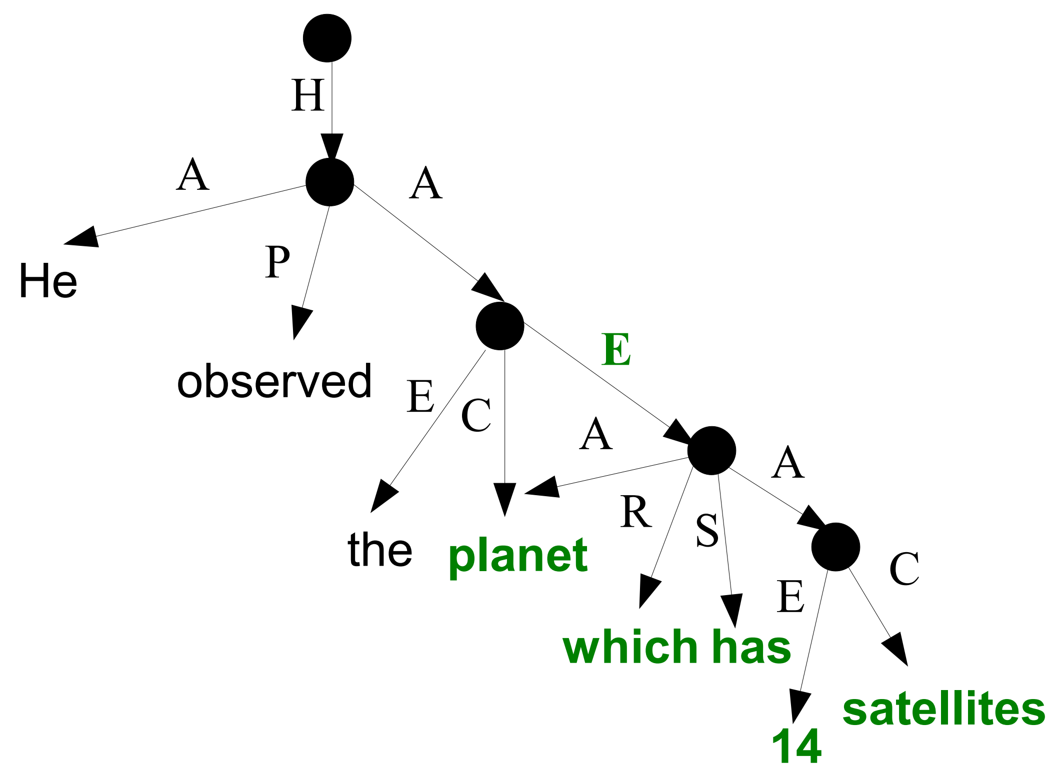
Center (C) Linker (L)

Elaborator (E) Relator (R)

The Semantic Structures

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

- A Scene can provide additional information on an established entity: it is then an **Elaborator Scene.**



Parallel Scene (H)

Participant (A)   Process (P)   State (S)

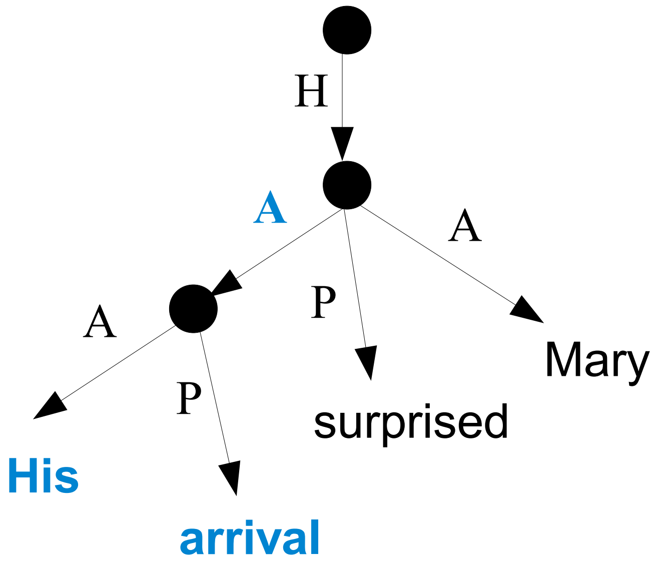
Center (C)   Elaborator (E)   Relator (R)

The Semantic Structures

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

- A Scene may also be a Participant in another Scene:

  It is then a **Participant Scene**.



Parallel Scene (H) Linker (L)

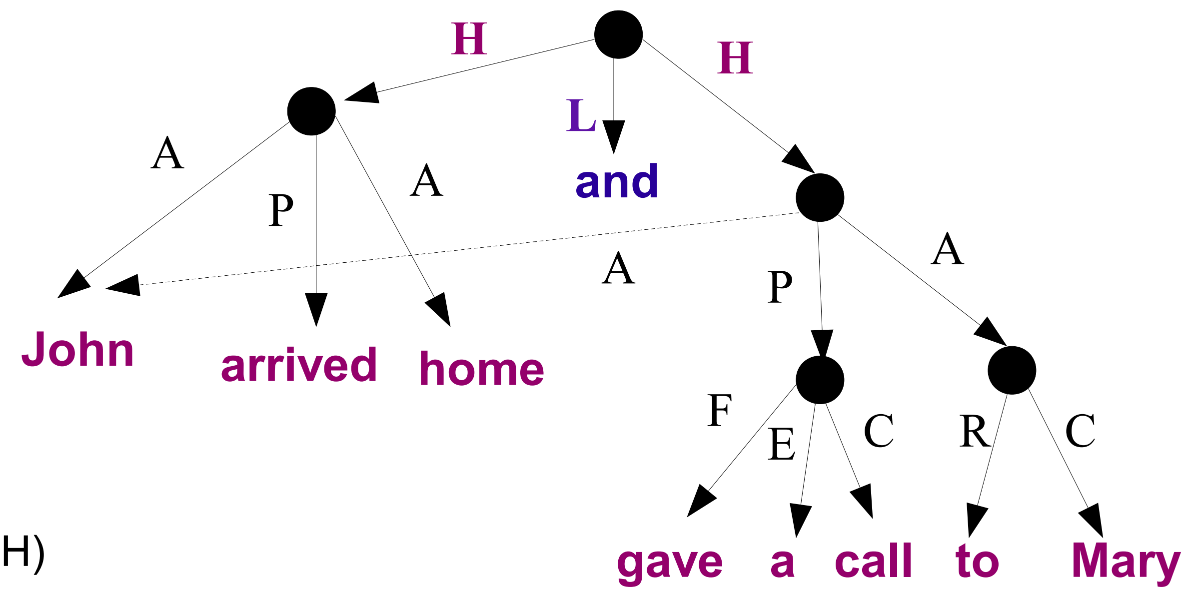
Participant (A) Process (P)

The Semantic Structures

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

- In the other cases, Scenes are annotated as **Parallel Scenes**.

  A **Linker** may be included.



Process (P) Function (F)

Participant (A) Parallel Scene (H)

Center (C) Linker (L)

Elaborator (E) Relator (R)

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.

SAMSA Computation

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

$[\textbf{John}]_A\ [\text{gave}_F]_{P-}\ [\textbf{Mary}]_A\ [\text{a}_E\ \textbf{call}_C]_{-P(CONT.)}$

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

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

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

Par_k - **Minimal center** of the kth Participant

$$Score_{Sen}(u) = \begin{cases} 1 & u \text{ is aligned to a word in } Sen \\ 0 & \text{otherwise} \end{cases}$$

SAMSA Computation

- Average over the input Scenes

- Non-splitting penalty: $\frac{n_{out}}{n_{inp}}$

Number of output sentences

Number of input Scenes

➡ We also experiment with SAMSA$_{abl}$, without non-splitting penalty.

SAMSA Computation

Example:

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

John arrived home. John called Mary. (output)

SAMSA Computation

Example:

(Sc1) John arrived home. (Sc2) John gave Mary a call. (Scene input)

(S1) John arrived home. (S2) John called Mary. (output sentences)

→ Scene and sentence identification

SAMSA Computation

Example:

(Sc1) John arrived home. (Sc2) John gave Mary a call. (input Scenes)

(S1) John arrived home. (S2) John called Mary. (output sentences)

Sc1 is matched to S1.

Sc2 is matched to S2.

➡ Scene-to-Sentence Matching

SAMSA Computation

Example:

(Sc1) [John]$_A$ [arrived]$_P$ [home]$_A$. (Sc2) John gave Mary a call.

(S1) John arrived home. (S2) John called Mary.

$$Score_{S1}(Sc\,1)=\frac{1}{2}\left(Score_{S1}(MR)+\frac{1}{K}\sum_{i=1}^{K} Score_{S1}(Par_k)\right)$$

$$Score_{S1}(Sc\,1)=\frac{1}{2}\left(1+\frac{1}{2}(1+1)\right)=1$$

➡ Scene scoring

SAMSA Computation

Example:

(Sc1) John arrived home. (Sc2) [John]$_A$ [gave$_F$]$_P$- [Mary]$_A$ [a$_E$ call$_C$]$_P$.

(S1) John arrived home. (S2) John called Mary.

$$Score_{S2}(Sc\,2)=\frac{1}{2}\left(Score_{S2}(MR)+\frac{1}{K}\sum_{i=1}^{K} Score_{S2}(Par_k)\right)$$

$$Score_{S2}(Sc\,2)=\frac{1}{2}\left(1+\frac{1}{2}(1+1)\right)=1$$

➡ Scene scoring

SAMSA Computation

Example:

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

John arrived home. John called Mary. (output)

$$SAMSA(input, output) = 1$$

$$SAMSA(input, output) = (\frac{2}{2}) \times (\frac{1}{2}) \times (Score_{S1}(Sc1) + Score_{S2}(Sc2))$$

$$SAMSA(input, output) = 1$$

➡ Average and 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):

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?

- 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):

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?

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

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

Correlation with Human Evaluation

	Reference-less				Reference-based		
	SAMSA Semi-Aut.	SAMSA Aut.	SAMSA_{abl} Semi-Aut.	SAMSA_{abl} Aut.	BLEU	SARI	Sent. with Splits
G	0.54	0.37	0.14	0.14	0.09	-0.77	0.09
P	-0.09	-0.37	**0.54**	**0.54**	0.37	-0.14	-0.49
S	0.54	0.71	-0.71	-0.71	-0.60	-0.43	0.83
AvgHuman	**0.58**	0.35	0.09	0.09	0.06	-0.81	0.14

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** – Strucutral Simplicity

- ➡ **SAMSA** obtained the **best correlation for AvgHuman.**
- ➡ **SAMSA_{abl}** obtained the **best correlation for Meaning Preservation.**

Correlation with Human Evaluation

	Reference-less				Reference-based		
	SAMSA Semi-Aut.	SAMSA Aut.	$\text{SAMSA}_{\text{abl}}$ Semi-Aut.	$\text{SAMSA}_{\text{abl}}$ Aut.	BLEU	SARI	Sent. with Splits
G	0.54	0.37	0.14	0.14	0.09	-0.77	0.09
P	-0.09	-0.37	0.54	0.54	0.37	-0.14	-0.49
S	**0.54**	**0.71**	-0.71	-0.71	-0.60	-0.43	0.83
AvgHuman	0.58	0.35	0.09	0.09	0.06	-0.81	0.14

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** – Strucutral Simplicity

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

	Reference-less				Reference-based		
	SAMSA Semi-Aut.	SAMSA Aut.	$\text{SAMSA}_{\text{abl}}$ Semi-Aut.	$\text{SAMSA}_{\text{abl}}$ Aut.	BLEU	SARI	Sent. with Splits
G	0.54	0.37	0.14	0.14	0.09	-0.77	0.09
P	-0.09	-0.37	**0.54**	**0.54**	0.37	-0.14	-0.49
S	0.54	0.71	-0.71	-0.71	-0.60	-0.43	0.83
AvgHuman	0.58	0.35	0.09	0.09	0.06	-0.81	0.14

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** – Strucutral Simplicity

➡ High similarity between the **Semi-Automatic** and **the Automatic** implementations. For **$\text{SAMSA}_{\text{abl}}$**, the ranking is the same.

Correlation with Human Evaluation

	Reference-less				Reference-based		
	SAMSA Semi-Aut.	SAMSA Aut.	SAMSA_{abl} Semi-Aut.	SAMSA_{abl} Aut.	BLEU	SARI	Sent. with Splits
G	0.54	0.37	0.14	0.14	0.09	-0.77	0.09
P	-0.09	-0.37	0.54	0.54	0.37	-0.14	-0.49
S	0.54	0.71	-0.71	-0.71	-0.60	-0.43	0.83
AvgHuman	0.58	0.35	0.09	0.09	0.06	-0.81	0.14

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** – Strucutral 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 3rd and 4th (**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.

- SAMSA could be also useful in a **more targeted Text Simplification**.

Thank you

Elior Sulem

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

Second Benchmark: https://github.com/eliorsulem/simplification-acl2018

eliors@cs.huji.ac.il

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

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM