האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

BLEU is Not Suitable for the Evaluation of Text Simplification

Elior Sulem, Omri Abend and Ari Rappoport
The Rachel and Selim Benin School of Computer Science and Engineering
The Hebrew University of Jerusalem

EMNLP 2018

BLEU

- BLEU (Paninени et al., 2002)
- Reference-based evaluation metric for MT
- Widely used in monolingual translation tasks, in particular:

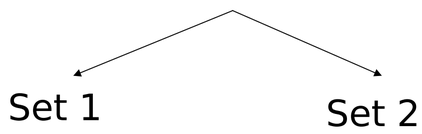
Text Simplification and **Split and Rephrase** Sub-task

HSplit Corpus

New Dataset

Input: Test set of Xu et al., 2016 (359 sentences)

Output: **Gold-standard sentence splitting**
Each sentence is modified by 4 annotators, according to **2 guideline sets**.



Split the original as much as possible, while preserving grammaticality, fluency and meaning

Split the original as much as possible, while preserving grammaticality, fluency and meaning, **if it simplifies the original**

- **4 structural paraphrases** for each of the sentences
- Average: **2.02** splits per sentence
  70 % of the sentences are split
- The mention of simplicity less affects the number of splits than the inter-annotator variability.
- It enriches the set of existed references focused on lexical operations (Xu et al., 2016) and is a new out-of-domain test set for Split and Rephrase.

Correlation with Human Evaluation

- **Hsplit as Reference Setting**

	Grammaticality (G)	Meaning Preservation (M)	Simplicity (S)	Structural Simplicity (StS)
BLEU	0.36	0.43	0.17	0.17
iBLEU	0.32	0.40	0.15	0.15
SARI	-0.05	-0.11	0.18	0.19
$-LD_{SC}$	**0.65**	**0.66**	**0.21**	**0.20**

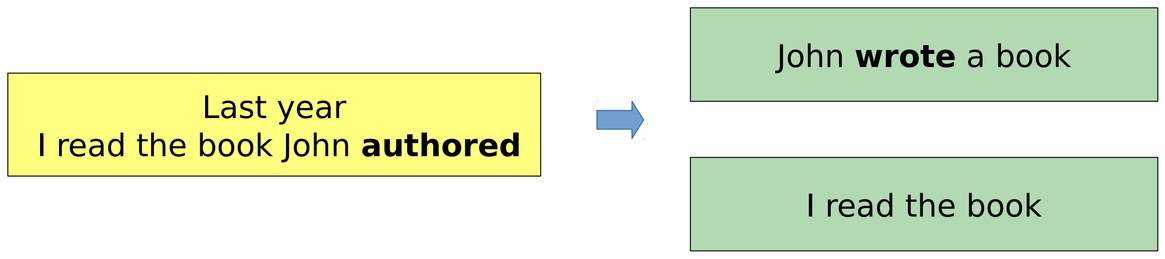
Spearman correlation at the sentence level between the automatic metrics and of human judgments

Systems: DSS, DSS^m, SEMoses, $SEMoses^m$, $SEMoses_{LM}$, $SEMoses^m{}_{LM}$
(Sulem et al., ACL 2018)

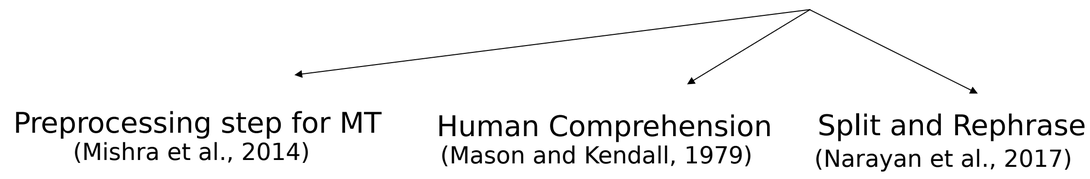
Correlation at the System-level: high for G (0.57), low for M (0.11), negative for S (-0.70) and StS (-0.60).

→ With references adapted to sentence splitting, BLEU still fails to assess this operation.

Text Simplification



Operations:
Word or phrase substitution, deletion, **sentence splitting**



Correlation with Human Evaluation

- **Standard Reference Setting**

	Systems/Corpora without Splits ; All Systems/Corpora			
	Grammaticality (G)	Meaning Preservation (M)	Simplicity (S)	Structural Simplicity (StS)
BLEU-1ref	0.43 ; 0.11	1.00 ; 0.08	-0.81 ; -0.60	-0.43 ; -0.67
BLEU-8ref	0.61 ; 0.26	0.89 ; 0.13	-0.59 ; -0.42	-0.11 ; -0.50
iBLEU-1ref	0.21 ; 0.02	0.93 ; 0.07	-0.85 ; -0.61	-0.61 ; -0.71
iBLEU-8ref	0.61 ; 0.26	0.89 ; 0.13	-0.59 ; -0.42	-0.11 ; -0.50
-FK	-0.21 ; -0.05	-0.57 ; -0.03	0.67 ; 0.51	0.39 ; 0.64
SARI-8ref	-0.64 ; -0.6	-0.86 ; -0.62	0.52 ; 0.26	0.00 ; -0.02
$-LD_{SC}$	0.29 ; 0.21	0.86 ; 0.51	-0.88 ; -0.68	-0.57 ; -0.52

Spearman correlation at the system level between the automatic metrics and of human judgments

- **Metrics**: BLEU, iBLEU (Sun and Zhou, 2012),Flesh Kincaid Grade Level (FK; Kincaid et al., 1975), SARI (Xu et al., 2016), Levenshtein distance to the source (LD_{SC}).
- **References:** 1ref: reference from Simple Wikipedia
  8ref: 8 crowdsourced references (Xu et al., 2016).
- **Human evaluation**: Sulem et al., ACL 2018, extended to Hsplit using the same protocol. We focus on the first 70 sentences for each system/corpus.
- **Systems/Corpora without Splits**: NTS (Nisioi et al., 2017) in 4 variants: h1, h4, w2vh1, w2vh4; Moses (Koehn et al., 2007); SBMT-SARI (Xu et al., 2016); Identity.
- **All Systems/Corpora**: Additionally includes the 4 HSplit corpora and the Hsplit average scores.

→ In all cases BLEU and iBLEU negatively correlate with Simplicity and Structural Simplicity.

→ Where sentence splitting is involved, the correlation with G and M disappears.

→ In this case, BLEU's correlation with M is considerably lower than that of $-LD_{SC}$ and its correlation with G is comparable.

→ Sentence-level correlation: for G and M the correlation with BLEU is lower than its correlation with $-LD_{SC}$ in both cases.

Conclusion

- Our findings suggest that BLEU should not be used for the evaluation of Text Simplification in general and sentence splitting in particular.
- It motivates the development of alternative methods for the evaluation of structural simplification. (Sulem et al., NAACL 2018).