Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification

Tomoyuki Kajiwara
Tokyo Metropolitan University
kajiwara-tomoyuki@ed.tmu.ac.jp

Atsushi Fujita
National Institute of Information and Communications Technology
atsushi.fujita@nict.go.jp

Quality Estimation for Text Simplification

- Data
  - Training: 505 sentence pairs
  - Test: 126 sentence pairs
- Four different evaluation criteria
  - Grammatically
  - Meaning preservation
  - Simplicity
  - Overall quality
- 3-class judgments for each criterion
  - {good, ok, bad}
- Evaluation metrics
  - A: Accuracy
  - E: Mean Absolute Error
  - F: Weighted F-score
- Best systems in QATS workshop
  - SimpleNets: neural networks
  - SMH: MT metrics
  - http://qats2016.github.io/

Motivation

- Neural networks are rather unstable because of the difficulty of training on a limited amount of data.
- MT metrics are incapable of properly capturing deletions and paraphrases that are prevalent in text simplification.

→ **In order to properly account for the surface-level inequivalency occurring in text simplification, we examine semantic similarity features based on word embeddings and paraphrase lexicons.**

Semantic Features Based on Word Alignments

1. Additive Embeddings Similarity

$$\mathrm{AES}(x,y) = \cos\left(\sum_{i=1}^{|x|} \vec{x}_i, \sum_{j=1}^{|y|} \vec{y}_j\right)$$

2. Average Alignment Similarity

$$\mathrm{AAS}(x,y) = \frac{1}{|x||y|} \sum_{i=1}^{|x|} \sum_{j=1}^{|y|} \cos(\vec{x}_i, \vec{y}_j)$$

3. Maximum Alignment Similarity

$$\mathrm{MAS}(x,y) = \frac{1}{|x|} \sum_{i=1}^{|x|} \max_j \cos(\vec{x}_i, \vec{y}_j)$$

4. Hungarian Alignment Similarity

$$\mathrm{HAS}(x,y) = \frac{1}{|\mathcal{H}|} \sum_{(i,j)\in\mathcal{H}} \cos(\vec{x}_i, \vec{y}_j)$$

5. Word Mover's Distance

$$\mathrm{WMD}(x,y) = \min \sum_{u=1}^{n} \sum_{v=1}^{n} \mathcal{A}_{uv} \mathrm{eud}(\vec{x}_u, \vec{y}_v)$$

6. Difference of Word Embeddings

$$\mathrm{DWE}(x,y) = \frac{1}{|x|} \sum_{i=1}^{|x|} \vec{x}_i - \frac{1}{|y|} \sum_{j=1}^{|y|} \vec{y}_j$$

7. Paraphrase Alignment Similarity

$$\mathrm{PAS}(x,y) = \frac{\mathrm{PA}(x,y) + \mathrm{PA}(y,x)}{|x| + |y|}$$

$$\mathrm{PA}(x,y) = \sum_{i=1}^{|x|} \begin{cases} 1 & \exists j : x_i \Leftrightarrow y_j \in y \\ 0 & \text{otherwise} \end{cases}$$

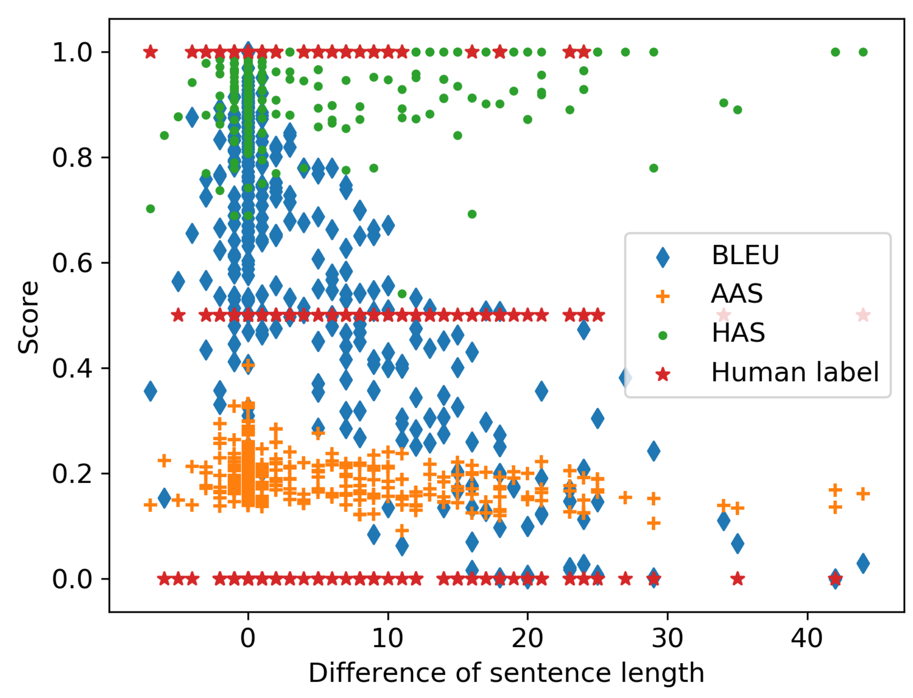
Evaluation using QATS dataset

- Classifiers based on our features greatly outperformed the state-of-the-art methods in terms of **Simplicity (Random Forest Classifier)** and **Overall quality (SVM Classifier)**.
- MT-baseline features do not help ours further.

→ **Word embeddings are superior to surface-level processing in finding corresponding words.**

System	Grammaticality			Meaning			Simplicity			Overall		
	A↑	E↓	F↑	A↑	E↓	F↑	A↑	E↓	F↑	A↑	E↓	F↑
Majority-class	**76.2**	18.3	65.9	57.9	29.0	42.5	55.6	29.4	39.7	43.7	28.2	26.5
Best score on QATS-2016 (Štajner+ 2016)	**76.2**	**17.1**	**71.8**	**69.1**	**20.2**	**68.1**	57.1	25.0	56.4	52.4	25.8	48.6
SVM Classifiers	MT-baseline: BLEU, METEOR, TER, WER											
MT-baseline	**76.2**	18.3	65.9	**66.7**	**20.2**	62.7	50.8	**26.2**	**48.3**	38.1	41.7	37.5
Our SVM	**76.2**	18.3	65.9	65.1	22.2	58.3	**57.1**	27.8	43.9	**57.9**	**23.4**	**57.7**
Our SVM w/ MT-baseline	**76.2**	18.3	65.9	**66.7**	21.0	**63.7**	**57.1**	27.0	46.9	47.6	29.0	46.8
Neural Network Classifiers	SimpleNets-MLP: multi-layer perceptron based on language model features											
SimpleNets-MLP (Paetzold and Specia, 2016)	**74.6**	**17.1**	**68.8**	**65.9**	**21.0**	**63.5**	53.2	27.0	49.8	38.1	32.5	33.7
Our MLP	68.3	24.6	66.9	59.5	25.4	56.4	**59.5**	**23.4**	**58.2**	**52.4**	**25.8**	**51.9**
Our MLP w/ MT-baseline	63.5	26.6	63.8	64.3	21.4	62.7	52.4	26.2	53.2	46.0	31.8	45.5
Random Forest Classifiers	SMH: based on automatic evaluation metrics and QE features for MT											
SMH-RandForest (Štajner+ 2016)	75.4	**17.5**	**71.8**	65.9	**20.6**	**64.4**	52.4	27.8	53.0	44.4	31.8	44.5
Our RandForest	**76.2**	18.3	65.9	**66.7**	23.0	63.2	**63.5**	**21.8**	**59.8**	**51.6**	**26.6**	**48.3**
Our RandForest w/ MT-baseline	**76.2**	18.3	65.9	61.9	24.6	57.6	62.7	22.6	56.1	46.0	29.0	43.6

Ablation on Accuracy	G	M	S	O
ALL	76.2	65.1	57.1	57.9
-AES	76.2	65.1	57.1	57.1
-MAS (Orig, Simp)	76.2	57.9	55.6	56.4
-MAS (Simp, Orig)	76.2	64.3	57.1	54.8
-PAS	76.2	57.9	55.6	53.2
-DWE	76.2	57.9	55.6	51.6
-WMD	76.2	57.9	55.6	46.8
-AAS	76.2	57.9	55.6	45.2
-HAS	76.2	57.9	55.6	35.7



Correlation	length	label
BLEU	-0.765	0.245
METEOR	-0.617	0.257
WMD	0.788	-0.215
AAS	-0.335	**0.318**
HAS	**0.061**	-0.050

HAS was not biased by the length difference almost at all, and **AAS** and highly correlated with the manually-labeled quality.

Example: A sentence pair judged "good" in terms of overall quality. **HAS reaches 0.85, while BLEU is 0.54.**

Original: While historians concur that the result itself was not manipulated, the voting process was neither free nor secret.

Simple: Most historians agree that the result was not fixed, but the voting process was neither free nor secret.

Hungarian Alignment