

MoverScore: A Novel Metric for Text Generation Evaluation

Wei Zhao



TECHNISCHE
UNIVERSITÄT
DARMSTADT



This is a recent work from the NLLG group, published at EMNLP2019.
(<https://www.aclweb.org/anthology/D19-1053.pdf>)

zhao@aiphes.tu-darmstadt.de



Text Generation: Machine Translation



TECHNISCHE
UNIVERSITÄT
DARMSTADT

English – detected ▼


↔



German ▼



The Oktoberfest
is the world's
largest Volksfest

×

Das Oktoberfest ist
das weltweit größte
Volksfest



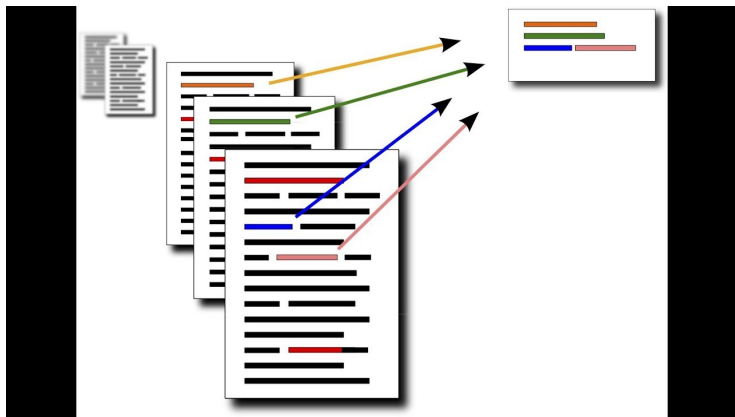
Open in Google Translate

Feedback

Text Generation: Text Summarization



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Text Generation Evaluation



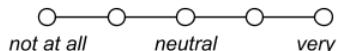
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Input: Bud Powell était un pianiste de légende.

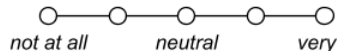
Reference: Bud Powell was a legendary pianist.

Candidate: Bud Powell was a great pianist.

How fluent is the sentence?



Does it accurately convey the meaning of the reference?



What metrics are used? (1)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

BLEU (Papineni et al. 2002): a metric in **Machine Translation**:

- ▶ A precision metric to measure **word overlap** between the system translation s and reference translation r
- ▶ Compute **n -gram precisions p_n** , along with different weights w_n
- ▶ Add brevity **penalty** to penalize short system translations

$$BLEU = \min(1, \frac{\text{len}(s)}{\text{len}(r)}) \exp \sum_{n=1}^N w_n \log p_n$$

What metrics are used? (2)



ROUGE (Lin et al. 2004): a metric in **Text Summarization**:

- ▶ A recall metric to measure word overlap between the system summary s and a set of reference summaries $\{r_1, \dots, r_M\}$

$$\text{ROUGE-N} = \frac{\sum_{i=1}^M \sum_{n=1}^N f(s, r_i)}{\sum_{i=1}^M \sum_{n=1}^N g(r_i)}$$

- ▶ $f(\cdot)$ counts the **n -gram words** matched between one summary s and reference r_i
- ▶ $g(\cdot)$ counts the n -gram words in one reference r_i
- ▶ Other variants like ROUGE-LCS and **ROUGE-SkipGram**

What metrics are used? (3)

More metrics:

- ▶ METEOR (Banerjee and Lavie 2005) accounts for precision and recall, with synonym matching for **Machine Translation**
- ▶ CIDEr (Vedantam et al., 2015) computes the cosine similarity between two n -gram vectors, the system and reference caption, for **Image Captioning**
- ▶ ROUGE-WE (Ng and Abrecht 2015) was proposed for **Text Summarization**
- ▶ Others like SPICE (Anderson et al., 2016) and chrF++ (Popović 2017)



- ▶ Neural systems increase the discrepancy between system and reference texts (Li et al., 2016, See et al., 2017)

System: Obama speaks to the media in Illinois.

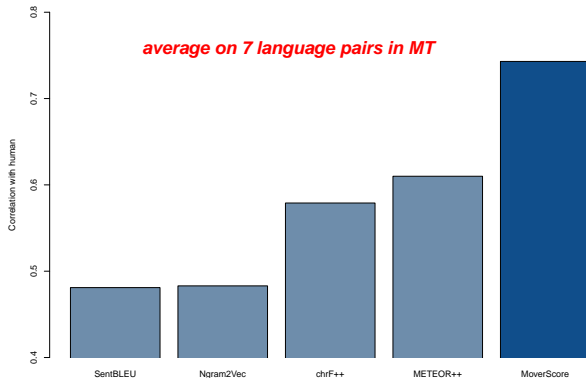
Reference: The president greets the press in Chicago.

Challenges for metrics (2)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Previous metrics show poor correlation with human judgments



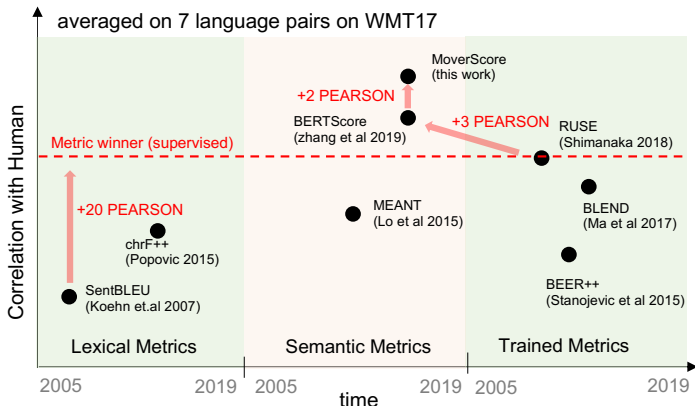
- ▶ Beyond BLEU (Wieting et al., 2019)
- ▶ BERTScore (Zhang et al., 2019)
- ▶ SemBLEU (Song et al., 2019)
- ▶ Sentence Mover's Similarity (Clark et al., 2019)
- ▶ Putting Evaluation in Context (Mathur et al., 2019)
- ▶ **MoverScore** (this work)

This is an active field, free to reach out if you are interested in doing a thesis in this topic :)

Evolution in metrics in Machine Translation



TECHNISCHE
UNIVERSITÄT
DARMSTADT



MoverScore: a versatile metric for text generation evaluation



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Instead of task-specific metrics, our metric:

- ▶ is versatile and general (applicable on multiple text generation tasks)
- ▶ has high correlation with human judgments

An example for computing MoverScore



TECHNISCHE
UNIVERSITÄT
DARMSTADT

System y : **Obama speaks** to the **media** in **Illinois**.

Reference y^* : The **president greets** the **press** in **Chicago**.

MoverScore is a set-based similarity metric to measure the semantic distance between two sequences of words, via summing up the transportation cost.

An example for computing MoverScore

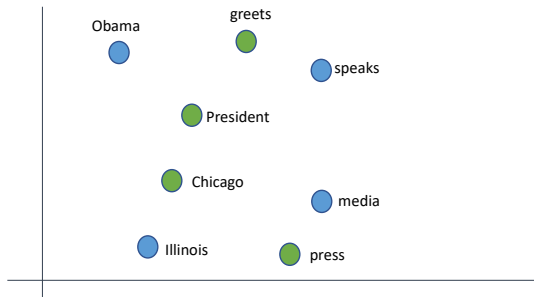


Abbildung: Step 1: obtain embeddings of y and y^*

An example for computing MoverScore



TECHNISCHE
UNIVERSITÄT
DARMSTADT

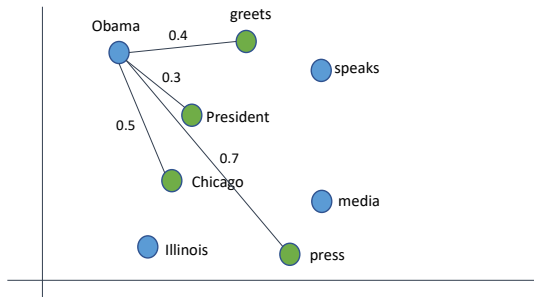


Abbildung: Step 2: get Euclidean distance of word embeddings

An example for computing MoverScore



TECHNISCHE
UNIVERSITÄT
DARMSTADT

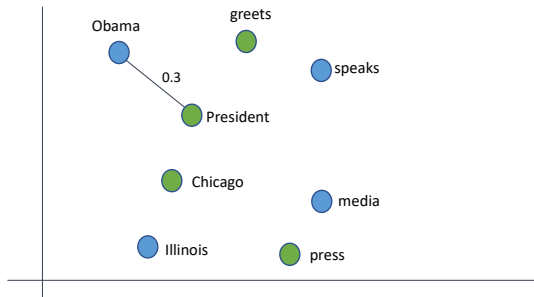


Abbildung: Step 3: find the shortest path to transport words

An example for computing MoverScore

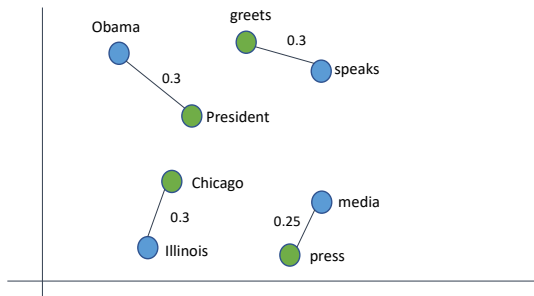


Abbildung: Step 3: find the shortest path to transport words

$$\text{score} = 0.3 + 0.3 + 0.3 + 0.25 = 1.25$$

An example for computing MoverScore

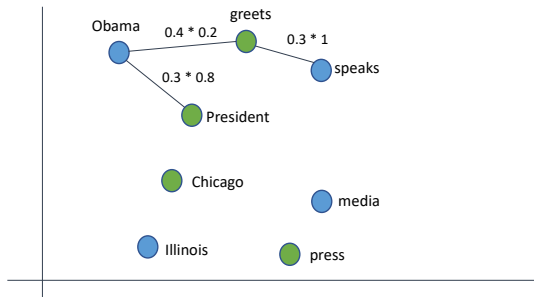


Abbildung: Step 3: find the optimal path to transport words

An example for computing MoverScore

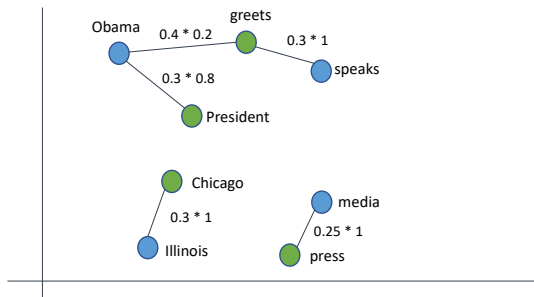


Abbildung: Step 3: find the optimal path to transport words

$$\text{metric} = 0.4 * 0.2 + 0.3 * 0.8 + 0.3 * 1 + 0.3 * 1 + 0.25 * 1 = 1.17$$

Summary of MoverScore



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Two sequences of words in system \mathbf{y} and reference text \mathbf{y}^*
- ▶ Obtain the embeddings of each word in \mathbf{y} and \mathbf{y}^*
- ▶ A distance matrix C denotes the distances of arbitrary word pairs in \mathbf{y} and \mathbf{y}^*
- ▶ Sum up how much cost the words travel from one set to other

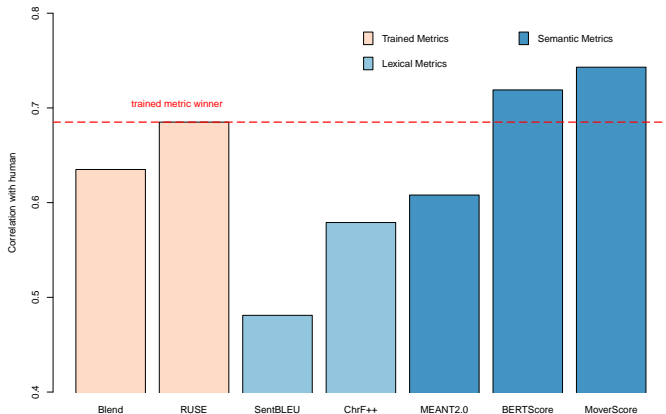
$$\text{WMD}(\mathbf{y}, \mathbf{y}^*) := \min_{F \in \mathbb{R}^{|\mathbf{y}| \times |\mathbf{y}^*|}} \sum C_{ij} \cdot F_{ij}$$

Results in Machine Translation



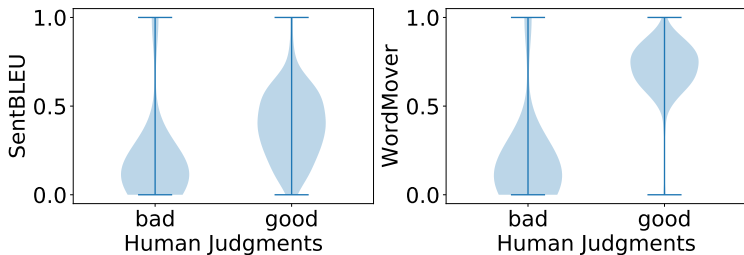
TECHNISCHE
UNIVERSITÄT
DARMSTADT

How do metrics correlate with human judgments on average?



Results in Machine Translation

How do metrics judge system translations with qualities at two extreme polar?

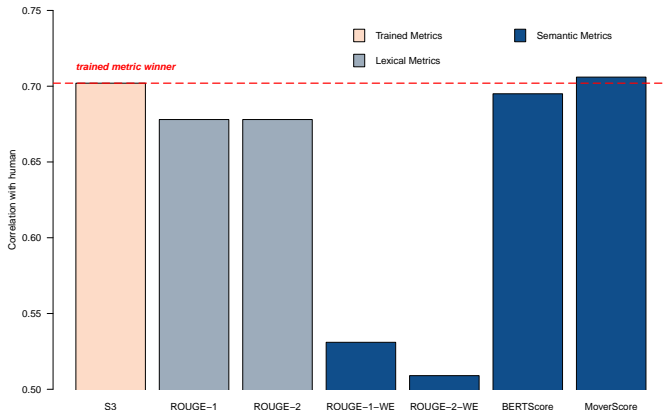


SentBLEU correctly assigns lower scores to low-quality translations but struggles in judging high-quality translations.

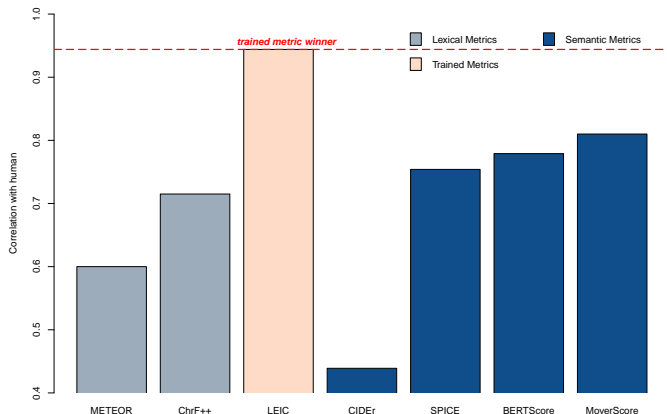
Results in Text Summarization



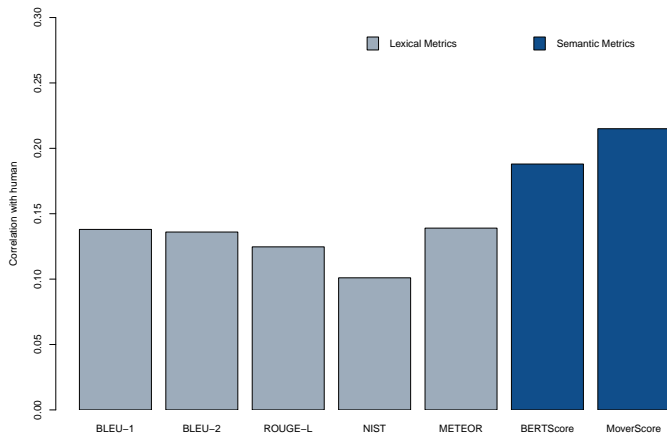
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Results in Image Captioning



Results in QA



Conclusion

- ▶ MoverScore is a new metric combining contextualized embedding and earth mover distance
- ▶ It achieves high correlation with human judgments on multiple NLG tasks.
- ▶ It clearly distinguishes low-quality from high-quality system-generated texts.

Code available in <https://github.com/AIPHES/emnlp19-moverscore>



Thänk yoü!