Predicting Machine Translation Adequacy from Document Embeddings Regression

Mihaela Vela and Liling Tan
Universität des Saarlandes
m.vela@mx.uni-saarland.de, liling.tan@uni-saarland.de

UNIVERSITÄT DES SAARLANDES
Marie Curie Actions

Introduction

Source (DE):
Auch der Aussenhandel bremste die Konjunktur.

Phrase-based MT:
The foreign trade braked the economy.

Neural MT:
External trade also slowed the economy.

Reference (EN):
Foreign goods trade had slowed, too.

- MT metric should penalize *poor lexical choice*, e.g. *braked*, and reward *semantically similar translations*, e.g. *external trade*
- We propose a *semantically grounded metric* using *Semantic Textual Similarity* (Agirre et al. 2014) regression model trained on document embeddings similarities (ZWICKEL + COMET)

Approach

Training

- Train a document embedding model

$$v(doc) = \frac{\sum_i^n v(w_i)}{n}; \quad doc = \{w_1, ..., w_n\} \quad (1)$$

- Calculate cosine similarity between *hyp* and *ref*

$$sim(hyp, ref) = v(hyp) \cdot v(ref) \quad (2)$$

- Compute the METEOR between *hyp* and *ref*
- Train a regressor where $X = \{sim(hyp_i, ref_i)\}$ and $Y = \{METEOR(hyp_i, ref_i)\}$
- N is the Gaussian prior, w and α are parameters estimated from the data.

$$p(y|X, w, \alpha) = N(y|X, w, \alpha) \quad (3)$$

Testing

- When testing, calculate $sim(hyp_{test}, ref_{test})$ and predict the METEOR score with the regressor (ZWICKEL)
- **Hack**: If METEOR scores 1.0 or 0.0 use METEOR instead of the regressor outputs to handles regression outliers (COMET)

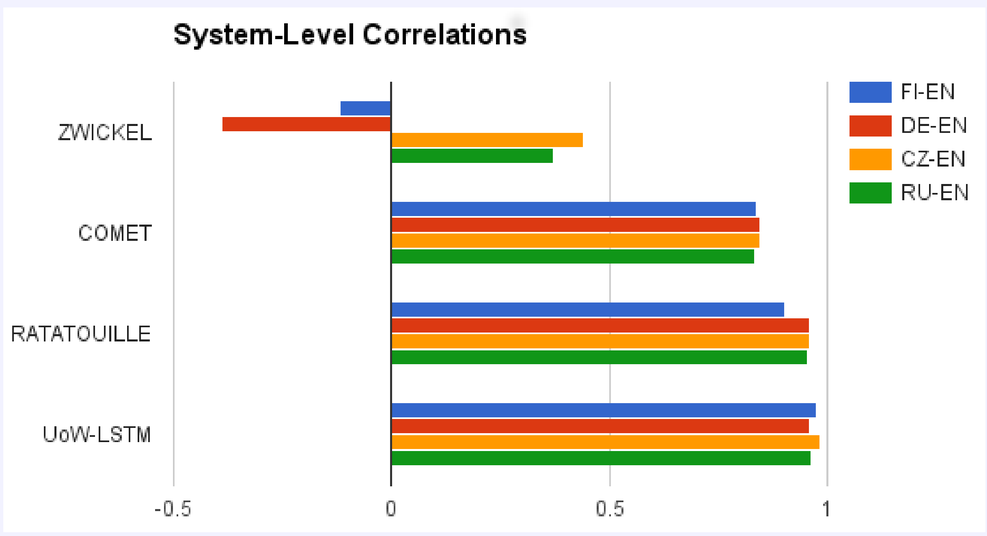
Results @ WMT15

Language pair / Averages	ZWICKEL	COMET
Finnish-English	-0.119 ± 0.037	0.836 ± 0.021
German-English	-0.389 ± 0.034	0.844 ± 0.020
Czech-English	0.441 ± 0.016	0.844 ± 0.010
Russian-English	0.371 ± 0.035	0.825 ± 0.021
Average	0.076 ± 0.031	0.819 ± 0.028
Pre-Trueskill Avg.	0.076 ± 0.031	0.837 ± 0.018
Spearman's Avg.	0.038 ± 0.076	0.718 ± 0.049

Table 1: Our System-level Correlations

Language pair / Averages	RATATOUILLE	UoW-LSTM
Finnish-English	0.902 ± 0.016	0.976 ± 0.008
German-English	0.958 ± 0.011	0.960 ± 0.010
Czech-English	0.961 ± 0.005	0.983 ± 0.003
Russian-English	0.955 ± 0.011	0.963 ± 0.009
Average	0.952 ± 0.010	0.976 ± 0.007
Pre-Trueskill Avg.	0.956 ± 0.014	0.976 ± 0.011
Spearman's Avg.	0.919 ± 0.039	0.916 ± 0.038

Table 2: Best System-level Correlations



Conclusion

- An experiment to find language independent alternatives to semantically grounded METEOR-like metrics

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. In SemEval 2014.

Milos Stanojevic, Amir Kamran, and Ondrej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In WMT15.

Benjamin Marie and Marianna Apidianaki. 2015. Alignment-based sense selection in meteor and the ratatouille recipe. In WMT15.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. Machine translation evaluation using recurrent neural networks. In WMT15.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471. This poster presentation is made possible with the funding from Marie Curie Alumni Association Micro-Travel Grant.