MT Quality Estimation for e-Commerce: Automatically Assessing the Quality of Machine Translation for Item Titles

Tsz Kin Lam*, José GC de Souza†, Nicola Ueffing†

*Department of Computational Linguistics,
Heidelberg University, Germany

†Machine Translation Science Team, Aachen, Germany

August 19 2019

Introduction

At ebay,

- Items transacted across countries with huge language diversity.
- Machine Translation (MT) system supporting translation of item titles.

However,

- The MT model is trained by rather clean sentences pairs.
- Titles of input language are noisy

$$\Longrightarrow \text{ Quality Estimation (QE) system}$$

- To filter poor translations out of the training set.
- To use the translations on the live site.
- To route the translations to post-editors.

Examples of Titles

Relatively well translated titles:

SRC: New XS Extreme Sport Sunglasses With Plastic Frames For Men And Women .
MT : NEU XS Extreme Sport Sonnenbrille mit Kunststoffrahmen für Männer und Frauen .

Noisy and poorly translated titles:

SRC: Handmade Duck Duct Tape Flower Pen - Set of $num * * YOU CHOOSE COLORS * *
MT : Handmade Ente Isolierband Blume Pen - Set $num * * Auswahl Farben * *

SRC: Seraph of the end こんにちわ Mikaela Yuichiro Hyakuya Arcylic Keychain Bag Pandent
MT: Seraph des Ende こんにちわ Mikaela Yuichiro hyakuya Arcylic SchlüsselAnhänger Tasche Pandent

Goal

To experiment two types of QE systems on our titles data:

(1) Predictor-Estimator (Kim et al. 2017):

- takes advantages of bitext, or called paired sentences, to pre-train the model.
- achieved state-of-the-art results on the WMT 2017 shared task.

(2) Siamese networks:

- does not require pre-training using bitext.
- is a popular metric-learning based method in the computer vision community.

Predictor-Estimator

- Predictor that can be pre-trained on bi-text for QE features prediction.
- Estimator, e.g., logistic regression, that takes the predicted QE features and outputs a confidence score.

Please refer the paper for details.

Siamese Network (In General)

1. Two information, e.g. a source sentence and a MT output, are processed individually but are compared at some point(s) of the network.
2. Distance-like metrics are used for comparison.
3. Wide variety of features extractors are possible, e.g., RNN and CNN.

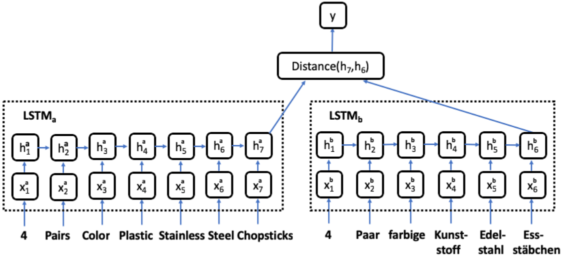

Figure: An illustration of siamese networks in MTQE, source: Ueffing et al. 2018.

Siamese Network: Training and Inference

Given data $\{(\text{src, mt, label})\}_{i=1}^{N}$, siamese networks minimizes contrastive loss (Hadsell et al. 2006):

$$\text{Loss} = \frac{y}{2}(1 - D)^2 + \frac{(1 - y)}{2}\max(D, 0.)^2 \tag{1}$$

where $D \in [0, 1]$ and $y \in \{0, 1\}$ are the predicted distance and label respectively. The label is defined as 0 if the mt is "GOOD" and vice versa. In inference,

$$\text{Label} = \begin{cases} \text{GOOD}, & \text{if } D \leq 0.5 \\ \text{BAD}, & \text{otherwise} \end{cases}$$

Our Siamese Networks

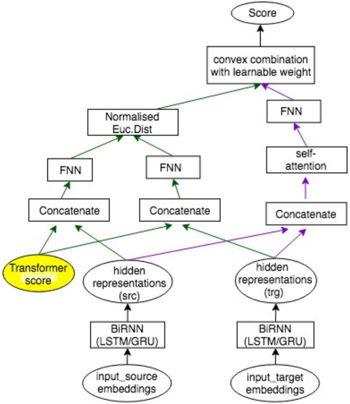

Figure: Architecture of our proposed Siamese network. The green arrows represents the major components of a siamese network used in Ueffing et al. 2018 whereas the purple arrows represents the components for the self-attention model. The transformer scores are added before FNNs.

Experiments - Data

Two experiments with translation direction from English to German:

- WMT 2017 sentence-level QE data
- Our in-house titles

Data	Purpose	train / dev / test	Avg. words per sent (en/de)
Europarl	pre-training	1.8M / 3k / N.A.	26.9 / 25.6
WMT	QE	23k / 1k / 2k	16.8 / 17.7
In-house bi-text	pre-training	287k / 3k / N.A.	14.3 / 13.7
In-house titles	QE	92k / 3k / 3k	12.4 / 11.7

Table: Corpus statistics for WMT and e-Commerce.

Results on WMT Test Dataset

System	Layer(s)	Pearson	MAE	RMSE	F1-weighted	F1-Good	F1-Bad
Predictor-Estimator †	Bi-GRU 500	0.4737	0.1304	0.1679	0.686	0.786	0.489
Siamese	Bi-GRU 250-250	-	-	-	0.675	0.766	0.493

Table: Comparison between Predictor-Estimator and Siamese on test data of WMT 2017

Remark:

1. † We used the DeepQuest (Ive et al. 2018) implementation of Predictor-Estimator to generate the result.
2. † Model is pre-trained on Europarl for 2 epochs
3. † F-1 scores are obtained by conversion of the HTER scores using a threshold of 0.3

Results on Machine Translated Titles

Model	Pre-trained	F1-Weighted	F1-Good	F1-Bad
DeepQuest	Europarl	61.5	44.2	74.1
DeepQuest	in-house bi-text	71.3	64.3	76.4
Siamese (NormEucDist)	NA	71.1	62.3	77.2
+Transformer NMT score	NA	72.6	65.8	77.3
Self-attention	NA	72	67.4	75.2
+Transformer NMT score	NA	73.3	68.2	76.9
Siamese (Convex)	NA	72.4	63.9	78.4
+Transformer NMT score	NA	76.0	70.2	80.0

Table: F1-scores on machine translated titles, En-De MT. All results are averaged over 3 runs.

Conclusions

We developed and evaluated methods for automatically assessing the quality of machine translated e-Commerce titles. Our siamese networks has:

1. comparable performance than Predictor-Estimator.
2. no need of gathering cleaned bi-text in related domain.
3. faster training speed.
4. about 3% gain improvement by adding transformer score as additional feature.

Thanks for coming !
Q & A

References

- Hadsell, Raia, Sumit Chopra, and Yann LeCun (2006). "Dimensionality reduction by learning an invariant mapping". In: *null*. IEEE, pp. 1735–1742.
- Ive, Julia, Frédéric Blain, and Lucia Specia (2018). "DeepQuest: a framework for neural-based quality estimation". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3146–3157.
- Kim, Hyun et al. (2017). "Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 17.1, p. 3.
- Ueffing, Nicola, José GC de Souza, and Gregor Leusch (2018). "Quality Estimation for Automatically Generated Titles of eCommerce Browse Pages". In: *Proc. NAACL-HLT 2018 (Industry Papers)*, pp. 52–59. DOI: 10.18653/v1/N18-3007.