Alternative Objective Functions for Training MT Evaluation Metrics

Miloš Stanojević
ILLC
University of Amsterdam
m.stanojevic@uva.nl

Khalil Sima'an
ILLC
University of Amsterdam
k.simaan@uva.nl

Abstract

MT evaluation metrics are tested for correlation with human judgments either at the sentence- or the corpus-level. *Trained* metrics ignore corpus-level judgments and are trained for high sentence-level correlation only. We show that training only for one objective (sentence or corpus level), can not only harm the performance on the other objective, but it can also be suboptimal for the objective being optimized. To this end we present a metric trained for corpus-level and show empirical comparison against a metric trained for sentence-level exemplifying how their performance may vary per language pair, type and level of judgment. Subsequently we propose a model trained to optimize *both objectives* simultaneously and show that it is far more stable than–and *on average* outperforms–both models on both objectives.

1 Introduction

Ever since BLEU (Papineni et al., 2002) many proposals for an improved automatic evaluation metric for Machine Translation (MT) have been made. Some proposals use additional information for extracting quality indicators, like paraphrasing (Denkowski and Lavie, 2011), syntactic trees (Liu and Gildea, 2005; Stanojević and Sima'an, 2015) or shallow semantics (Rios et al., 2011; Lo et al., 2012) etc. Whereas others use different matching strategies, like n-grams (Papineni et al., 2002), treelets (Liu and Gildea, 2005) and skip-bigrams (Lin and Och, 2004). Most metrics use several indicators of translation quality which are often combined in a linear model whose weights are estimated on a training set of human judgments.

Because the most widely available type of human judgments are relative ranking (RR) judgments, the main machine learning method used for training the metrics were based on the learning-to-rank framework (Li, 2011). While the effectiveness of this framework for training evaluation metrics has been confirmed many times, e.g., (Ye et al., 2007; Duh, 2008; Stanojević and Sima'an, 2014; Ma et al., 2016), so far there is no prior work exploring alternative objective functions for training learning-to-rank models. Without exception, all existing learning-to-rank models are trained to rank sentences while completely ignoring the corpora judgments, likely because human judgments come in the form of sentence rankings.

It might seem that sentence and corpus level tasks are very similar but that is not the case. Empirically it has been shown that many metrics that perform well on the sentence level do not perform well on the corpus level and vice versa. By training to rank sentences the model does not necessarily learn to give scores that are well scaled, but only to give higher scores to better translations. Training for the corpus level score would force the metric to give well scaled scores on the sentence level.

Human judgments of sentences can be aggregated in different ways to hypothesize human judgments of full corpora. However, this fact has not been used so far to train learning-to-rank models that are good for ranking different corpora.

This work fills-in this gap by exploring the merits of different objective functions that take corpus level judgments into consideration. We first create a learning-to-rank model for ranking corpora and compare it to the standard learning-to-rank model that is trained for ranking sentences. This comparison shows that performance of these two objectives can vary radically depending on the chosen meta-evaluation method. To tackle this prob-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 20–25
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2004

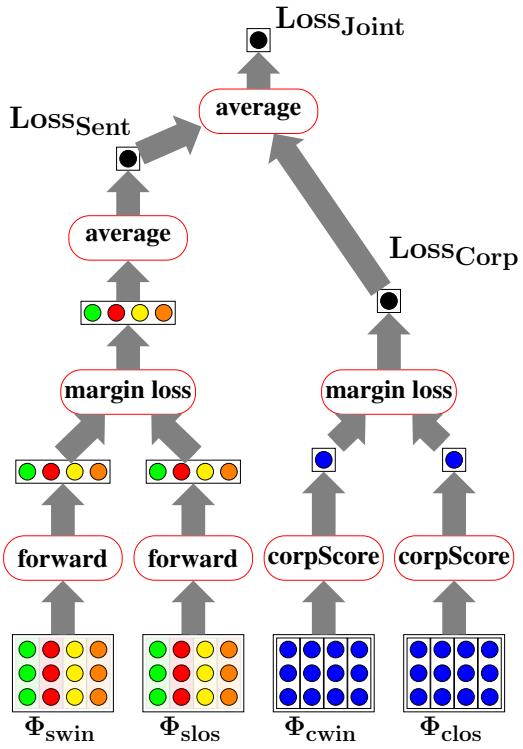

Figure 1: Computation Graph

lem we contribute a new objective function, inspired by multi-task learning, in which we train for both objectives simultaneously. This multi-objective model behaves a lot more stable over all methods of meta-evaluation and achieves a higher correlation than both single objective models.

2 Models

All the models that we define have one basic function in common, we call it a $forward(\cdot)$ function, that maps the features of any sentence to a single real number. That function can be any differentiable function including multi-layer neural networks as in (Ma et al., 2016), but here we will stick with the standard linear model:

$$forward(\phi) = \phi^T \mathbf{w} + b$$

Here ϕ is a vector with feature values of a sentence, $\mathbf{w}$ is a weight vector and b is a bias term. Usually in training we would like to process a mini-batch of feature vectors $\mathbf{\Phi}$, where $\mathbf{\Phi}$ is a matrix in which each column is a feature vector of individual sentence in the mini-batch or in the corpus. By using broadcasting we can rewrite the previous definition of the $forward(\cdot)$ function as:

$$forward(\mathbf{\Phi}) = \mathbf{\Phi}^T \mathbf{w} + b$$

Now we can define the score of a sentence as a sigmoid function applied over the output of the $forward(\cdot)$ function because we want to get a score between 0 and 1:

$$sentScore(\phi) = \sigma(forward(\phi))$$

As the corpus level score we will use just the average of sentence level scores:

$$corpScore(\Phi) = \frac{1}{m} \sum sentScore(\Phi)$$

where m is the number of sentences in the corpus.

Next we present several objective functions that are illustrated by the computation graph in Figure 1.

2.1 Training for Sentence Level Accuracy

Here we use the training objective very similar to BEER (Stanojević and Sima'an, 2014) which is a learning-to-rank framework that finds a separating hyper-plane between "good" and "bad" translations. Unlike BEER, we use a max-margin objective instead of logistic regression.

For each mini-batch we randomly select m human relative ranking pairwise judgments and after extracting features for all the sentences taking part in these judgments we put features in two matrices Φ_{swin} and Φ_{slos}. These matrices are structured in such a way that for judgment i the column i in Φ_{swin} contains the features of the "good" translation in the judgment and the column i in Φ_{slos} the features of the "bad" translation.

We would like to maximize the average margin that would separate sentence level scores of pairs of translations in each judgment. Because the squashing sigmoid function does not influence the ranking we can directly optimize on the unsquashed forward pass and require that the margin between "good" and "bad" translation is at least 1:

$$\Delta_{sent} = forward(\Phi_{swin}) - forward(\Phi_{slos})$$

$$Loss_{Sent} = \frac{1}{m} \sum max(0, 1 - \Delta_{sent})$$

2.2 Training for Corpus Level Accuracy

At the corpus level we would like to do a similar thing as on the sentence level: maximize the distance between the scores of "good" and "bad" corpora. In this case we have additional information that is not present on the sentence level: we know not only which corpus is (according to humans) better, but also by *how much* it is better. For

that we can use one of the heuristics such as the Expected Wins (Koehn, 2012). We can use this information to guide the learning model by how much it should separate the scores of two corpora.

For doing this we use an approach similar to Max-Margin Markov Networks (Taskar et al., 2003) where for each training instance we dynamically scale the margin that should be enforced. We want the margin between the scores Δ_{corp} to be at least as big as the margin between the human scores Δ_{human} assigned to these systems. In one mini-batch we will use only a randomly chosen pair of corpora with feature matrices Φ_{cwin} and Φ_{clos} for which we have a human comparison. The corpus level loss function is given by:

$$\Delta_{corp} = corpScore(\Phi_{cwin}) - corpScore(\Phi_{clos})$$
$$Loss_{Corp} = max(0, \Delta_{human} - \Delta_{corp})$$

2.3 Training Jointly for Sentence and Corpus Level Accuracy

In this model we optimize both objectives jointly in the style of multi-task learning (Caruana, 1997). Here we employ the simplest approach of just tasking the interpolation of the previously introduced loss functions.

$$Loss_{Joint} = \alpha \cdot Loss_{Sent} + (1 - \alpha) \cdot Loss_{Corp}$$

The interpolation is controlled by the hyper-parameter α which could in principle be tuned for good performance, but here we just fix it to 0.5 to give both objectives equal importance.

2.4 Feature Functions

The feature functions that are used are reimplementation of many (but not all) feature functions of BEER. Because the point of this paper is about the exploration of different objective functions we did not try to experiment with more complex feature functions based on paraphrasing, function words or permutation trees.

We use just simple precision, recall and 3 types of F-score (with β parameters 1, 2 and 0.5) over different "pieces" of translation:

- character n-grams of orders 1,2,3,4 and 5
- word n-grams of orders 1,2,3 and 4
- skip-bigrams of maximum skip 2 and ∞ (similar to ROUGE-S2 and ROUGE-S* (Lin and Och, 2004))

One final feature deals with length-disbalance. If the length of the system and reference translation are a and b respectively then this feature is computed as $\frac{max(a,b)-min(a,b)}{min(a,b)}$. It is computed both for word and character length.

3 Experiments

Experiments are conducted on WMT13 (Macháček and Bojar, 2013), WMT14 (Machacek and Bojar, 2014) and WMT16 (Bojar et al., 2016) datasets which were used as training, validation and testing datasets respectively.

All of the models are implemented using TensorFlow[1] and trained with L2 regularization $\lambda = 0.001$ and ADAM optimizer with learning rate 0.001. The mini-batch size for sentence level judgments is 2000 and for the corpus level is one comparison. Each model is trained for 200 epochs out of which the one performing best on the validation set for the objective function being optimized is used during the test time.

We show the results for the relative ranking (RR) judgments correlation in Table 1. For all language pairs that are of the form *en-X* we show it under the column X and for all the language pairs that have English on the target side we present their average under the column *en*.

RR corpus vs. sentence objective The corpus-objective is better than the sentence-objective for both corpus and sentence level RR judgments on 5 out of 7 languages and also on average correlation.

RR joint vs. single-objectives Training for the joint objective improves even more on both levels of RR correlation and outperforms both single-objective models on average and on 4 out of 7 languages.

Making confident conclusions from these results is difficult because, to the best of our knowledge, there is no principled way of measuring statistical significance on the RR judgments. That is why we also tested on direct assessment (DA) judgments available from WMT16. On DA we can measure statistical significance on the sentence level using Williams test (Graham et al., 2015) and on the corpus level using combination of hybrid-supersampling and Williams test (Graham and Liu, 2016). The results of correlation with human judgment are for sentence and corpus level are shown in Table 2.

[1] https://www.tensorflow.org/

Objective	en	cs	de	fi	ro	ru	tr	Average
sent	**0.963**	0.977	0.737	0.938	0.922	0.905	0.937	0.912
corpus	0.944	0.982	**0.765**	0.940	0.917	**0.907**	**0.954**	0.916
joint	**0.963**	**0.983**	0.748	**0.951**	**0.933**	0.905	0.946	**0.918**

(a) Corpus level

Objective	en	cs	de	fi	ro	ru	tr	Average
sent	0.347	0.405	0.345	0.304	0.293	0.382	0.304	0.340
corpus	0.337	**0.414**	0.349	**0.307**	0.292	0.385	**0.325**	0.344
joint	**0.350**	0.410	**0.356**	0.296	**0.299**	**0.396**	0.312	**0.346**

(b) Sentence level

Table 1: Relative Ranking (RR) Correlation. The corpus level correlation is measured with Pearson r and sentence level with Kendall τ

Objective	en-ru	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Average
sent	$\mathbf{0.9113}^{C}_{J}$	0.9839^{C}	0.8483^{C}	$\mathbf{0.9556}^{C}_{J}$	0.8348^{C}	0.8888^{C}	$\mathbf{0.9706}^{C}_{J}$	0.9133
corpus	0.9086	0.9790	0.8032	0.9121	0.7933	0.8857	0.9011	0.8833
joint	0.9111^{C}	$\mathbf{0.9844}^{C}_{S}$	$\mathbf{0.8488}^{C}_{S}$	0.9545^{C}	$\mathbf{0.8399}^{C}_{S}$	$\mathbf{0.8935}^{C}_{S}$	0.9647^{C}	**0.9138**

(a) Corpus level

Objective	en-ru	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Average
sent	0.6655^{C}	0.6478^{C}	0.4930^{C}	0.4608^{C}	0.5066^{C}	0.5535^{C}	0.5800^{C}	0.5582
corpus	0.5632	0.5676	0.3913	0.3644	0.3771	0.4306	0.4579	0.4503
joint	0.6668^{C}	$\mathbf{0.6631}^{S}_{C}$	$\mathbf{0.5019}^{S}_{C}$	0.4608^{C}	$\mathbf{0.5276}^{S}_{C}$	0.5564^{C}	0.5830^{C}	**0.5657**

(b) Sentence level

Table 2: Direct Assessment (DA) Pearson r Correlation. Super- and sub-scripts S, C and J signify that the model outperforms with statistical significance ($p < 0.05$) the model trained for sentence, corpus or joint objective respectively. Bold marks that the system has outperformed both other models significantly.

DA corpus vs. other objectives On DA judgments the results for corpus level objective are completely different than on the RR judgments. On DA judgments the corpus-objective model is significantly outperformed on both levels and on all languages by both of the other objectives.

This shows that gambling on one objective function (being that sentence or corpus level objective) could give unpredictable results. This is precisely the motivation for creating the joint model with multi-objective training.

DA joint vs. single objectives By choosing to jointly optimize both objectives we get a much more stable model that performs well both on DA and RR judgments and on both levels of judgment. On the DA sentence level, the joint model was *not* outperformed by any other model and on 3 out of 7 language pairs it significantly outperforms both alternative objectives. On the corpus level results are a bit mixed, but still joint objective outperforms both other models on 4 out of 7 language pairs and also it gives higher correlation on average.

4 Conclusion

In this work we found that altering the objective function for training MT metrics can have radical effects on performance. Also the effects of the objective functions can sometimes be unexpected: the sentence objective might not be good for sentence level correlation (in case of RR judgments) and the corpus objective might not be good for corpus level correlation (in case of DA judgments). The difference among objectives is better explained by different types of human judgments: the corpus objective is better for RR while sentence objective is better for DA judgments.

Finally, the best results are achieved by training for both objectives at the same time. This gives

an evaluation metric that is far more stable in its performance over all methods of meta-evaluation.

Acknowledgments

This work is supported by NWO VICI grant nr. 277-89-002, DatAptor project STW grant nr. 12271 and QT21 project H2020 nr. 645452.

References

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 199–231. http://www.aclweb.org/anthology/W/W16/W16-2302.

Rich Caruana. 1997. Multitask learning. *Machine Learning* 28(1):41–75. https://doi.org/10.1023/A:1007379606734.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Kevin Duh. 2008. Ranking vs. Regression in Machine Translation Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '08, pages 191–194. http://dl.acm.org/citation.cfm?id=1626394.1626425.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, CA.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*. Denver, Colorado.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *Proceedings of International Workshop on Spoken Language Translation*. http://www.mt-archive.info/IWSLT-2012-Koehn.pdf.

Hang Li. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '04. https://doi.org/10.3115/1218955.1219032.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 25–32. http://www.aclweb.org/anthology/W/W05/W05-0904.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '12, pages 243–252. http://dl.acm.org/citation.cfm?id=2393015.2393048.

Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. 2016. Maxsd: A neural machine translation evaluation metric optimized by maximizing similarity distance. In Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng, editors, *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing and 24th International Conference on Computer Processing of Oriental Languages*. Springer International Publishing, Kunming, China, pages 153–161.

Matous Machacek and Ondrej Bojar. 2014. Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 293–301. http://www.aclweb.org/anthology/W/W14/W14-3336.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 45–51. http://www.aclweb.org/anthology/W13-2202.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. Tine: A metric to assess mt adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 116–122. http://www.aclweb.org/anthology/W11-2112.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 202–206. http://www.aclweb.org/anthology/D14-1025.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 396–401. http://aclweb.org/anthology/W15-3050.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2003. Max-Margin Markov Networks. In *NIPS 2014 - Advances in Neural Information Processing Systems 27*.

Yang Ye, Ming Zhou, and Chin-Yew Lin. 2007. Sentence Level Machine Translation Evaluation As a Ranking Problem: One Step Aside from BLEU. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '07, pages 240–247. http://dl.acm.org/citation.cfm?id=1626355.1626391.