PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 108 JUNE 2017 307–318

Comparative Quality Estimation for Machine Translation Observations on Machine Learning and Features

Eleftherios Avramidis

German Research Center for Artificial Intelligence (DFKI Berlin), Language Technology Lab

Abstract

A deeper analysis on Comparative Quality Estimation is presented by extending the state-of-the-art methods with adequacy and grammatical features from other Quality Estimation tasks. The previously used linear method, unable to cope with the augmented features, is replaced with a boosting classifier assisted by feature selection. The methods indicated show improved performance for 6 language pairs, when applied on the output from MT systems developed over 7 years. The improved models compete better with reference-aware metrics.

Notable conclusions are reached through the examination of the contribution of the features in the models, whereas it is possible to identify common MT errors that are captured by the features. Many grammatical/fluency features have a good contribution, few adequacy features have some contribution, whereas source complexity features are of no use. The importance of many fluency and adequacy features is language-specific.

1. Introduction

The need for automatically predicting the quality of Machine Translation (MT) output has lead into the development of Quality Estimation (QE; Specia et al., 2009). Whereas most QE tasks aim at a single judgment, there have been concerns on how confident one can be in quantifying quality. Humans seem to have difficulty in scoring the quality of translations, particularly in defining the distinction between the level of quality each score represents (Callison-Burch et al., 2007). A solution would be to reduce the requirements for the ground truth, by favouring ordinality against cardinality. This can be done by eliciting judgments of relative quality, through direct comparisons between two or more translation items (Duh, 2008). For problems that

Corresponding author: eleftherios.avramidis@gmail.com

Cite as: Eleftherios Avramidis. Comparative Quality Estimation for Machine Translation Observations on Machine Learning and Features. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 307–318.
doi: 10.1515/pralin-2017-0029.

require comparisons of performance, it may be beneficial to neglect qualitative observations that are irrelevant to the comparison and may interfere with the decision.

Following this idea, we are focusing on Comparative QE as the automatic process of analyzing two or more translations produced by various MT systems and employing machine learning (ML) to express a judgment about how they compare in terms of quality. Although a considerable amount of research has employed this concept for various applications, such as system combination, statistical MT tuning and evaluation, there has been little analysis of the very concept of Comparative QE per se.

In this paper we attempt to extend the relatively limited state-of-the-art work and investigate the factors that play an important role for the task. In particular we will:

- bring features from other QE tasks to Comparative QE: introduce adequacy features, augment the grammatical ones with CFG rules and position indicators,
- observe whether linear methods in this problem can cope with the amount and the type of the advanced features and suggest instead an ensemble classifier
- improve on previous work regarding the competition with reference-aware metrics, confirming that elaborate features and ML may provide more information about relative translation quality than the comparison with the references,
- show which quality indicators are important for comparing MT outputs by investigating their contribution in the produced models, identify the MT errors that make these features useful for the automatic comparison of the translations,
- use feature selection methods to select an optimal number of features in order to improve the performance of the learning method or to achieve the same performance with a smaller amount of features,
- indicate the importance of grammatical features and confirm that the contribution of specific grammatical features is language-specific
- empirically confirm that source complexity features are not useful for predicting a comparison between automatic translations.

2. Related Work

The concept of Comparative QE, although not explicitly defined, has been used in many MT related tasks. In particular, previous works perform it as they:

(a) predict a continuous score independently for each system output and then they rank the outputs based on their individual score (e.g. Specia et al., 2009),

(b) use binary classification or regression with a cut-off value, to accept/reject a basic system and then back-off to another system without judging it (Quirk, 2004),

(c) use binary classification to compare two systems (Yasuda et al., 2002) or

(d) use an ordinal ranking (Herbrich et al., 1999) to compare an undefined number of systems (Hopkins and May, 2011; Avramidis et al., 2011; Formiga et al., 2013).

In this paper we are going to follow on the latter work. It essentially extends the binary classification (b), with the difference that the underlying classifier is system-agnostic and that it decides on comparisons for all possible pairs. Contrary to the continuous regression approach, the ordinal model only learns a relative notion of the translation quality, by having quality indicators from all compared outputs.

Formiga et al. (2013) confirm that ordinal regression makes better predictions as compared to ordering MT outputs, based on separate regression models over absolute scores of adequacy. When it comes to learning from ordinal rankings, Avramidis and Popović (2013) set the state-of-the-art performance for German-English, in the frame of a WMT shared task in QE (Bojar et al., 2013),

Previous work has motivated the use of grammatical features focusing in specific structures (eg. Mutton et al., 2007), feature selection was motivated by Specia et al. (2009), whereas an analysis of features was done by Felice and Specia (2012); nevertheless all the above work is limited to non-Comparative QE.

As compared to previous work, here we extend the state-of-the-art on Comparative QE by increasing the human correlation through the use of a Gradient Boosting classifier. We add additional linguistically-informed features inspired from other tasks. We also present a detailed analysis of the contribution of (a) the individual features, (b) the feature selection and (c) the learning methods. Our models exceed all previous experiments in coverage, as they expand into 6 language directions and are learned on outputs from heterogeneous MT systems developed within a period of 7 years.

3. Methods

3.1. Problem definition

This work aims at developing an empirical system which is able to order multiple translation outputs in the same way humans would do. In particular, the system is given one source sentence and several translations which have been produced for this sentence. The goal is to *rank* them, i.e. to order the translations based on their quality after deriving several qualitative criteria over the translations.

We define a ranking $R = \{s, \mathbf{t}, \mathbf{r}\}$ where a source sentence s is associated with a set of translations $\mathbf{t} = (t_1, t_2, \ldots, t_m)$, as t_j is the j-th translation of s and m the number of the translations. Each set of translations $\mathbf{t}$ is associated with a list of ordinal judgments (ranks) $\mathbf{r} = (r_1, r_2, \ldots, r_n)$, where r_j is the judgment on translation t_j, as compared to the other translations in $\mathbf{t}$. This kind of qualitative ordering does not imply any absolute or generic measure of quality. Ranking takes place on a sentence level, which means that the inherent mechanism focuses on only one sentence at a time, considers the available translation options and makes a decision. Any assigned rank has therefore a meaning only for the sentence-in-focus and given the particular alternative translation candidates. Each source sentence $s^{(i)}$ is associated with a set of translations $\mathbf{t}^{(i)} = (t_1^{(i)}, t_2^{(i)}, \ldots, t_m^{(i)})$ where $t_j^{(i)}$ is the j-th translation of the i-th source sentence and m the number of the translations. Each list of translations is associated with a list containing relative judgments (ranks) $\mathbf{r}^{(i)} = (r_1^{(i)}, r_2^{(i)}, \ldots, r_n^{(i)})$ where $r_j^{(i)}$ is the judgment on the j-th translation of the i-th source sentence.

Counts: number of tokens and unknown words, number of occurrences of the target word within the target hypothesis (type/token ratio), number of commas and dots,
Parsing: PCFG parsing for both source and target side: the sentence log-likelihood, the number of n-best trees, the number of VPs in the best parse tree
Source complexity features: average source token length, average number of translations per source word in the sentence, percentage of unigrams/bigrams/trigrams in frequency quartiles 1 (lower frequency words) and 4 (high frequency words) in a corpus of the source language, percentage of source sentence unigrams seen in a corpus
Contrastive scoring: the METEOR score using the competing translations as references

Counts: avg. chars per word, count of nums and of tokens with non-alphabetic characters
Language model: smoothed probability from 3-gram and 5-gram LM, 3-gram perplexity
IBM Model 1: scores on both directions
Contrastive scoring: smoothed BLEU; precision, recall, frag. penalty of METEOR
Unknown words: first and last position of unknown words (absolute and normalized to the length of the sentence), average and standard dev. of the positions of unknown words
Rule-based correction: total errors, comma/parenthesis+space, uppercase sentence start

Table 1. Upper: Features for the baseline feature set. Lower: Features for the augmented feature set, added to the baseline features and the grammatical features of Section 3.2

A *feature vector* is defined as $\mathbf{x}^{(i)} = G(s^{(i)}, \mathbf{t}^{(i)})$ and it is created from every pair of source and its translations $(s^{(i)}, \mathbf{t}^{(i)})$, where $i = 1, 2, \ldots n$. The function G that produces the feature vector given a source and its translations is referred to as *feature generation*. Each feature vector $\mathbf{x}^{(i)}$ derived from the i-th source sentence and the corresponding list of ranks define an *instance* $I^{(i)} = (\mathbf{x}^{(i)}, \mathbf{r}^{(i)})$ and a *training set* of n instances is consequently defined as $T = \{(\mathbf{x}^{(i)}, \mathbf{r}^{(i)})\}_{i=1}^{n}$. A *ranker* is a function which given a feature vector $\mathbf{x}^{(i)}$ produces a list of *predicted* ranks $\widehat{\mathbf{r}}^{(i)}$. The goal of the *learning process* is therefore given the training set T to define a ranker that minimizes the total error between the predicted list of ranks and the golden list of ranks: $\sum_{i=1}^{m} \mathcal{E}(\mathbf{r}^{(i)}, \widehat{\mathbf{r}}^{(i)})$.

3.2. Feature generation

The **baseline feature set** (upper Table 1) consists of features that had the optimal performance as reported in previous work, i.e. the baseline and the best performing ranking QE features of WMT (Bojar et al., 2013). The **augmented feature set** extends the baseline set with features from non-Comparative QE (lower Table 1). Additionally more fluency features are added, as deemed helpful in the baseline, and adequacy features are introduced, as they were absent. These features are described below:

We count the **node labels of the parse tree**, namely NPs, VPs, PPs, verbs, nouns and for every node label we get the minimum, maximum and average depth/height of its positions in the tree and the average and standard deviation of its position. Every parse tree is decomposed into **Context-Free Grammar (CFG) rules** and for every rule, we get the number of occurrences and statistics about its height and depth in the tree. For the rules that contain a VP or a verb, two additional features indicate their distance from the beginning and the end of the sentences. This is of particular interest for translations into German, where the position of the VPs in the sentence is important.

A set of **alignment features** is produced as the nodes between the source and the target trees are aligned based on the scores of the lexical IBM-1 model (Zhechev, 2009). For every node alignment, we get the count of the aligned nodes in the sentences, the count of occurrences of the target CFG rules whose heads are aligned to the similar rules in the source, the depth of the source node in the source tree and the distance of the aligned nodes (if related to verbs) from the beginning and the end of the sentence.

This process got all possible alignments of node labels, resulting into 154,657 features. Nevertheless, many of these features are sparse, since they depend on the appearance of grammatical phenomena, so we used some sparsity heuristics resulting into 139 features: the monolingual CFG features including VPs and NPs with more than 20k occurrences (5+5 features), CFG alignment features including VPs with more than 10k occurrences (5) and NPs with more than 30k occurrences (5), CFG position features with more than 24k occurrences (5), rule-based corrections with more than 1k occurrences (4) and from the rest of the features, the ones with more than 51k occurrences (110 features). This selection aims at making the experiments computationally feasible, although there is no evidence that the reduced set is optimal.

3.3. Learning Methods and Evaluation

The ranker performs pairwise classification (Avramidis and Popović, 2013). The baseline uses Logistic Regression with the Newton-Raphson algorithm including Stepwise Feature Set Selection. As an advanced method, after preliminary experiments[1], we chose a Gradient Boosting of 100 decision trees and 100 boosting stages, limiting the maximum depth of the individual estimators to 3 and presorting data in order to find splits faster. Feature selection is done with Recursive Feature Elimination with cross-validation (RFECV) using SVM (Herbrich et al., 1999) with a linear kernel.

The predicted ranking is evaluated based on its correlation with human rankings, using Cross Validation with 10 folds over the entire dataset. The correlation metric is Kendall's tau as per WMT12: ties and cases of equal disagreement are removed from the test sets, whereas predicted ties are counted as discordant pairs, occasionally leading to negative taus.[2] Significance tests are based on the theoretical two-tailed t-

[1] including Decision Trees, Gaussian Naïve Bayes, kNN, LDA, Log. Regression with L2 Regularisation, Adaboost, Bagging, ExtRa Trees, and Random Forest. The boosting was tested with both 50 and 100 trees

[2] The evaluation setup differs from that of Bojar et al. (2013) to allow more robust testing, so here we rerun and evaluate their best methods as our baseline. Under our evaluation setup they result into slightly different scores

test of tau and confidence intervals by bootstrap resampling ($n = 1000$, $\alpha = 0.05$). NDCG is considered as an additional ranking metric (Järvelin and Kekäläinen, 2002).

4. Experiments

The experiments are performed on MT output from WMT annotated with human rankings (WMT2008-2014; e.g. Bojar et al., 2013) for English to German, French, Spanish and vice-versa, but advanced feature engineering is done only for German due to the increased MT errors for this language. A separate model is trained for every language direction. Per language pair, there are about 7k sentences from the news domain translated by about 100 systems. Translations of each sentence are grouped randomly into batches of 5 and ranked by various annotators. This provides 13k-25k batches, resulting into 64k to 100k pairwise comparisons. The vast majority of the systems are phrase-based and variations, whereas only 5% are rule-based.

Feature generation and learning are run with Qualitative (Avramidis, 2016), PCFG is run with the Berkeley Parser pre-trained on the TIGER, TueBaD/Z, AncoRa and FTB treebanks (Petrov et al., 2006) and rule-based correction is run with LanguageTool[3].

4.1. Ranking performance

In this experiment (a) we test whether the predicted rankings have any correlation with human rankings, (b) we compare the augmented ranking mechanism against the baseline and a random ranking and (c) we compare the augmented ranking mechanism against state-of-the-art reference-aware metrics. The metrics compared are: BLEU with sentence-level smoothing (Papineni et al., 2001), METEOR, (Denkowski and Lavie, 2014), rgbF (Popović, 2012), WER and TER (Snover et al., 2006).

Results The results (Table 2) indicate that (a) the predicted rankings have significant correlation with human rankings with a t-test p-value almost zero, (b) the predicted rankings are significantly better than random ones. The augmented ranking mechanism has achieved improved correlation against the baseline ranking mechanism.

A notable improvement over the baseline is that (c) the augmented ranking mechanism performs significantly better than the state-of-the-art reference-aware automatic metrics on a sentence level for the language pairs involving German, where focused feature engineering took place. It also outperforms other metrics in language pairs where the feature engineering from other language pairs was adopted, apart from one metric, METEOR, which is on par with the ranking mechanism. This confirms that elaborate features and ML may provide more information about relative translation quality than direct comparison with references.

[3] http://languagetool.org

lang.	basel.	augm.	random	BLEU	METEOR	rgbF	TER	WER
de-en	0.26*	0.28*	-0.14	-0.22 ‡	0.23 ‡	0.16 ‡	-0.02 ‡	0.15 ‡
en-de	0.15*	0.17*	-0.17	-0.42 ‡	0.13 ‡	0.10 ‡	-0.09 ‡	-0.15 ‡
es-en	0.11*	0.22*	-0.18	-0.19 ‡	0.22 ◇	0.16 ‡	-0.02 ‡	0.13 ‡
en-es	0.11*	0.12*	-0.17	-0.21 ‡	0.12 ◇	0.09 ◇	-0.10 ‡	0.08 ‡
fr-en	0.18*	0.19*	-0.18	-0.18 ‡	0.20 ◇	0.15 ‡	-0.02 ‡	0.16 ‡
en-fr	0.20*	0.21*	-0.15	-0.12 ‡	0.18 ◇	0.15 ‡	-0.03 ‡	0.15 ‡

‡: augmented ranking mechanism is significantly better than metric
◇: augmented ranking mechanism is significantly as good as metric
*: correlation with humans is significant, with a measured $p < 4 \cdot 10^{-20}$

Table 2. Basic vs. augmented ranking mechanism with random ranking and automatic metrics, concerning correlation with human judgments (tau) on segment-level

4.2. Observations on the baseline features

Useful conclusions concerning the contributions of various features can be drawn by examining the estimated beta coefficients of the logistic regression model of the baseline. For every coefficient, the null hypothesis of it being equal to zero has been rejected with a χ-test. The sign (positive/negative) of the coefficient indicates whether the feature has a positive or a negative contribution to the selection of the translation by the humans. Also, since the feature values are normalized with their mean and variance, the coefficient may provide indications for the importance of the features on the final decision. Some observations on the beta coefficients (Table 3) are:

Number of unknown words: Although OOVs are not necessarily untranslated words, when two translations of the same source have a different amount of unknown words, it is more likely that the one with the most of them has failed to translate some.

Overall amount of tokens: Statistical systems often omit the translation of some source words. This occurs when words suggested by the translation model reduce dramatically the overall score during the decoding process. Manual evaluation indicates that this occurs with long-distance re-ordering of German verbs, not scored properly by the language model. Therefore, when a translation has less words than its competitor, it may be the case that a useful word was omitted. *Additional words* also occur as a translation error, e.g. when phrases chosen during the decoding of a phrase-based system overlap partially. A special case of this, when the same word is repeated in the generated translation (type/token ratio) is given a negative coefficient.

Contrastive scoring: When more than one systems perform the same translation, they often convey more correct information collectively than each of them. Therefore, a system output that agrees more with the majority of the other systems is more likely to be preferred as the best translation.

The number of verb phrases (VPs) is connected with the fluency, as a result of the parser having tried to analyze the sentence and identify the VPs. Among translation errors, it is more likely that a VP is not formed properly, than having superfluous VPs formed by mistake. Therefore, it is observed that if a translation has more VPs than its competitor, it is more likely to be chosen. Similarly, when the parser analyses a translation, it creates **n-best lists with trees** with all possible grammatical analyses. The size of the list can indicate how ambiguous the parse is and therefore a translation with fewer n-best trees is more preferable for comparing translations. The **parse log-likelihood** also has a positive contribution, as an indication of grammaticality.

Punctuation count indicates that translation systems often make mistakes with punctuation and it is more likely to select a translation when it has fewer commas, or when it has more dots. Systems erroneously create too many commas or omit dots.

Finally, there is little explanation of the low, albeit negative contribution of the **tri-gram LM probability**, since one would expect that a higher probability would be preferable. One could assume that this is interacting with some other features, e.g. to favour grammatical features over the LM, or that some MT systems overvalue the LM score, which is also the reason for the omission of German verbs, mentioned earlier.

There can also be conclusions about the features which were assigned a zero coefficient. Using this, we can see that out of the **non-comparative QE features** only the punctuation features, the type/token ratio and the tri-gram probability helped, added to the target sentence length, which already existed as a feature. **Source complexity features** have been also assigned zero coefficients, so we can confirm that they play no role in the comparison between translations and that they do not introduce any useful knowledge about the *relative ability* of the systems to translate these sentences.

4.3. Machine Learning method and Feature Selection

Here, we investigate (a) the effect of adding the augmented feature set on the baseline model with Logistic Regression (b) the possibility to reduce the amount of features by performing Feature Selection (c) the improvements by using an ensemble instead of a linear classifier and finally (d) the effect of adding/removing features.

Feature Selection is applied only for German-English and English-German on a sub-set of the full-dataset. Since RFECV does not scale well, it is run on a stratified sample resulting into the 2.5% of the original sentences of a single fold for German-English and the 5% for English-German[4]. The selected feature set was used to train and evaluate the ranking model with 10-folded cross-validation, as above.

Results The results of using RFECV and Gradient Boosting can be seen in Table 4. Simply adding the augmented feature set on the baseline model with the Logistic

[4] Although this small sample is not guaranteed to be enough for feature selection, we will show that it is enough for reducing the feature size without harming the overall performance

feature name (target sentence)	β
number of unknown words	-0.58
number of tokens	0.50
contrastive METEOR	0.29
number of VPs	0.17
number of n-best trees	-0.17
type/token radio	-0.14
number of commas	-0.11
sentence parse log-likelihood	0.08
3-gram probability	-0.05
number of dots	0.04
…other features of Table 1	0.00

Table 3. Logistic Regression coefficients for the baseline, in descending order of absolute values

lang.	method	set	tau	NDCG
de-en	LogReg	basic	0.261	0.730
		full	0.110	0.680
		RFECV	0.181	0.716
	GradBoost	basic	0.265	0.736
		full	**0.280**	**0.742**
		RFECV	**0.276**	**0.739**
en-de	LogReg	basic	0.151	0.725
		full	0.034	0.703
		RFECV	0.020	0.696
	GradBoost	basic	0.138	0.723
		full	**0.170**	**0.733**
		RFECV	**0.174**	**0.731**

Table 4. Performance of the basic, the full feature set and the result of the RFECV with Logistic Regression and Gradient Boosting

Regression causes a significant drop, indicating that this method is not capable of handling such an amount and type of features, possibly because it cannot handle non-linear indicators. RFECV improves significantly the performance of Logistic Regression on the augmented feature set for German-English, but it still does not reach the performance of the same algorithm with the baseline set. For English-German, both the full set and the RFECV lead to almost zero correlation.

When it comes to using the advanced feature set, Gradient Boosting achieves significantly better performance than Logistic Regression. Using RFECV to reduce the full set has a negligible effect on the model trained with Gradient Boosting. Although the usage of RFECV did not improve the performance, it is interesting that the number of features (139) was reduced to less than the half, but the correlation remained the same. Reducing the amount of features can be of interest in an application environment, since it also reduces the computation. The above observation can also be seen in Figure 1, which depicts the increase in the classification quality, as features are added in the model. The optimal set for German-English contains 41 features, whereas the English-German one contains 56 features. The performance reaches already high levels with an amount of about 25 features and after a few fluctuations it enters a plateau where more features do not have a significant implication to the model.

4.4. Observations on the advanced features

Whereas 139 features were passed to Feature Selection, the latter favoured a significantly smaller number of features, nevertheless leading to the same performance. We can use the results of the selection to (a) identify important differences between the baseline and the augmented set and (b) compare between the two language directions. Some observations on the selection (Table 5) are:

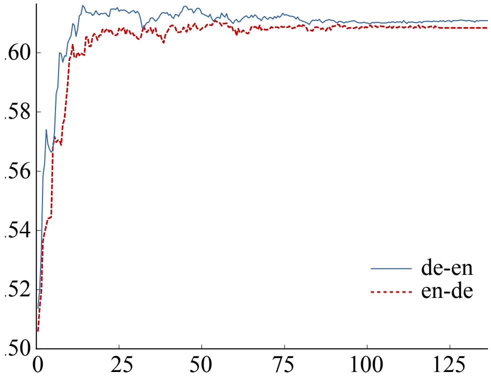

Figure 1. Number of features selected by RFECV vs. classification accuracy

language pair	de-en	en-de
Tree nodes		
nouns (count)		+
nouns (average position)	+	+
nouns (std of positions)	+	
NPs (count)	+	+
NPs (average position)	+	
NPs (std of positions)	+	+
VPs (std of positions)		+
VPs (avg, max tree height)	+	+
PPs (count, std of positions)		+
CFG rules		
NP→DT-NN (count)	+	
PP→IN-NP (count)	+	
VP→TO-VP (count)	+	
S→VP (position from end)		+
VP→VP (position from end)		+
Aligned CFG rules and nodes		
S→NP-VP (count/depth/pos.)	+	
NP (count)		+

Table 5. Grammatical features selected by RFECV

Augmented vs. baseline feature set: Although source complexity features were ruled out during Logistic Regression, Feature Selection for the augmented set favours few features that do include source information through the alignment of grammatical structures between source and target. For German-English, these are the statistics of the alignment of the simplest CFG sentence rule (S→NP-VP), whereas for English-German the aligned NPs. The contribution of these alignments is reasonable, given their grammatical operation and density. Additionally, this indicates that although simple features based on source information may be of little use, targeted features that capture translation adequacy on particular structures can still be of high relevance for comparing translations. Finally, it is worth noting that single features from the basic ranking mechanism have been replaced by a multitude of more specific features with similar functionality (e.g. the count of VPs has been replaced with counts of VPs within more fine-grained rules). This can be attributed to the advanced learning method which can handle better a larger amount of partially overlapping features.

Comparison between language pairs: Language-specific differences are shown by the grammatical that were automatically selected. The ones selected for English-German indicate the importance of the *position of the VPs and the PPs* in the sentence, obviously justified by the German positional requirements. This is in contrast to German-English, which get no features referring to the position of VPs or PPs. For the direction into English we can note the CFG rules that relate with grammatical phenomena which may be often mistranslated, such as the NPs with a determiner and a noun, the VPs containing a gerund and the PPs with the preposition "in".

5. Conclusion and further work

We have built on top of previous state-of-the-art work on Comparative Quality Estimation by introducing adequacy features and severely augmenting the grammatical/fluency features with CFG rules and position indicators. Logistic Regression used previously cannot handle properly the advanced features, possibly because they are non-linearly separable, so we introduced a Gradient Boosting classifier that could cope better with the problem and improve the performance of the ranking.

We tested the methods with 6 language directions by training on the output of systems spanning 7 years of development. The models can compete better against state-of-the-art reference-aware metrics on the segment-level, particularly when language-specific feature engineering took place, confirming previous observations that elaborate features with ML can compete direct scoring against references. The contribution of grammatical features is notable and it is possible to identify common MT errors that justify the empirically estimated contribution of particular indicators. The use of most grammatical features strongly depends on the target language, e.g. position of VPs is important for German. The majority of the features indicate fluency, few features indicate adequacy, whereas source complexity features are of no importance.

Although these experiments are based on empirical analysis on the output of a broad set of MT systems, we are aware that we are missing some significant representation of Neural MT, which has changed considerably the quality and the error types of MT. Investigations to this direction will be inevitably part of further work.

Acknowledgment This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 645452 (QT21)

Bibliography

Avramidis, Eleftherios. Qualitative: Python Tool for MT Quality Estimation Supporting Server Mode and Hybrid MT. *The Prague Bulletin of Mathematical Linguistics*, 106:147–158, 2016.

Avramidis, Eleftherios and Maja Popović. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 329–336, Sofia, Bulgaria, 2013.

Avramidis, Eleftherios, Maja Popović, David Vilar, and Aljoscha Burchardt. Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of WMT*, pages 65–70, Edinburgh, Scotland, 2011.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 12–58, Sofia, Bulgaria, 2013.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, 2007.

Denkowski, Michael and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, 2014.

Duh, Kevin. Ranking vs. regression in machine translation evaluation. *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, 2008.

Felice, Mariano and Lucia Specia. Linguistic Features for Quality Estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Canada, 2012.

Formiga, Lluís, Lluís Màrquez, and Jaume Pujantel. Real-life Translation Quality Estimation for MT System Selection. In *Proceedings of MT Summit XIV*, pages 69–76, Nice, France, 2013.

Herbrich, Ralf, Thore Graepel, and Klaus Obermayer. Support Vector Learning for Ordinal Regression. In *International Conference on Artificial Neural Networks*, pages 97–102, 1999.

Hopkins, Mark and Jonathan May. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, 2011.

Järvelin, Kalervo and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Mutton, Andrew, Mark Dras, Stephen Wan, and Robert Dale. GLEU: Automatic Evaluation of Sentence-Level Fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, 2007.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176(W0109-022), IBM, 2001.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proc. of ACL*, pages 433–440, Sydney, Australia, 2006.

Popović, Maja. rgbF: An Open Source Tool for n-gram Based Automatic Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 98(98):99–108, 2012.

Quirk, Chris. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of LREC2004*, volume 4, pages 825–828, Lisbon, Portugal, 2004.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. A Study of Translation Error Rate with Targeted Human Annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas*, 2006.

Specia, Lucia, M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Annual Meeting of the European Association for Machine Translation*, pages 28–35, Barcelona, Spain., 2009.

Yasuda, Keiji, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. Automatic machine translation selection scheme to output the best result. In *Proceedings of LREC2002*, pages 525–528, Las Palmas, Spain, 2002.

Zhechev, Ventsislav. Unsupervised Generation of Parallel Treebank through Sub-Tree Alignment. *Prague Bulletin of Mathematical Linguistics*, 91:89–98, 2009.

Address for correspondence:
Eleftherios Avramidis
eleftherios.avramidis@gmail.com
Alt Moabit 91c, 10559 Berlin, Germany