Exploring Consensus in Machine Translation for Quality Estimation

Carolina Scarton and **Lucia Specia**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{c.scarton,l.specia}@sheffield.ac.uk

Abstract

This paper presents the use of consensus among Machine Translation (MT) systems for the WMT14 Quality Estimation shared task. Consensus is explored here by comparing the MT system output against several alternative machine translations using standard evaluation metrics. Figures extracted from such metrics are used as features to complement baseline prediction models. The hypothesis is that knowing whether the translation of interest is similar or dissimilar to translations from multiple different MT systems can provide useful information regarding the quality of such a translation.

1 Introduction

While Machine Translation (MT) evaluation metrics can rely on the similarity of the MT system output to reference (human) translations as a proxy to quality assessment, this is not possible for MT systems in use, translating unseen texts. Quality Estimation (QE) metrics are used in such settings as a way of predicting translation quality. While reference translations are not available for QE, previous work has explored the so called *pseudo-references* (Soricut and Echihabi, 2010; Soricut et al., 2012; Soricut and Narsale, 2012; Shah et al., 2013). Pseudo-references are alternative translations produced by MT systems different from the system that we intend to predict quality for (Albrecht and Hwa, 2008). These can be used to provide additional features to train QE models. Such features are normally figures resulting from automatic metrics (such as BLEU, Papineni et al. (2002)) computed between pseudo-references and the output of the given MT system.

Soricut and Echihabi (2010) explore pseudo-references for document-level QE prediction to rank outputs from an MT system. The pseudo-references-based features are BLEU scores extracted by comparing the output of the MT system under investigation and the output of an off-the-shelf MT system, for both the target and the source texts. The statistical MT system training data is also used as pseudo-references to compute training data-based features. The use of pseudo-references has been shown to outperform strong baseline results. Soricut and Narsale (2012) propose a method that uses sentence-level prediction models for document-level QE. They also use a pseudo-references-based feature (based in BLEU) and claim that this feature is one of the most powerful in the framework.

For QE at sentence-level, Soricut et al. (2012) use BLEU based on pseudo-references combined with other features to build the best QE system of the WMT12 QE shared task.[1] Shah et al. (2013) use pseudo-references in the same way to extract a BLEU feature for sentence-level prediction. Feature analysis on a number of datasets showed that this feature contributed the most across all datasets.

Louis and Nenkova (2013) apply pseudo-references for summary evaluation. They use six systems classified as "best systems", "mediocre systems" or "worst systems" to make the comparison, with ROUGE (Lin and Och, 2004) as quality score. They also experiment with a combination of the "best systems" and the "worst systems". The use of only "best systems" led to the best results. Examples of "bad summaries" are said not to be very useful because a summary close to the worst systems outputs can mean that either it is bad or it is too different from the best systems outputs in terms of content. Albrecht and Hwa (2008) use pseudo-references to improve MT evaluation by combining them with a single human reference. They show that the use of pseudo-references im-

[1] http://www.statmt.org/wmt12/

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 342–347,
Baltimore, Maryland USA, June 26–27, 2014.

proves the correlation with human judgements.

Soricut and Echihabi (2010) claim that pseudo-references should be produced by systems as different as possible from the MT system of interest. This ensures that the similarities found among the systems' translations are not related to the similarities of the systems themselves. Therefore, the assumption that a translation from system *X* shares some characteristics with a translation from system *Y* is not a mere coincidence. Another way to make the most of pseudo-references is to use an MT system known as generally better (or worse) than the MT system of interest. In that case, the comparison will lead to whether the MT system of interest is similar to a good (or bad) MT system.

However, in most scenarios it is difficult to rely on the average translation quality of a given system as an absolute indicator of its quality. This is particularly true for sentence-level QE, where the quality of a given system can vary significantly across sentences. Finding translations from MT systems that are considerably different can also be a challenge. In this paper we exploit pseudo-references in a different way: measuring the consensus among different MT systems in the translations they produce. As sources of pseudo-references, we use translations given in a multi-translation dataset or those produced by the participants in the WMT translation task for the same data. While some MT systems can be similar to each other, for some language pairs, such as English-Spanish, a wide range of MT systems with different average qualities are available. Our hypothesis is that by using translations from several MT systems we can find **consensual information** (even if some of the systems are similar to the one of interest). The use of more than one MT system is expected to smooth out the effect of "coincidences" in the similarities between systems' translations.

This paper describes the use of consensual information for the WMT14 QE shared task (USHEFF-consensus system), simulating a scenario where we do not know the quality of the pseudo-references, nor the characteristics of any MT systems (the system of interest or the systems which generated the pseudo-references). We participated in all variants of Task 1, sentence-level QE, for both for scoring and ranking. Section 2 explains how we extracted consensual information for all tasks. Section 3 shows our official results compared to the baselines provided. Section 4 presents some conclusions.

2 Consensual information extraction

The consensual information is exploited in two different ways in Task 1. Task 1.1 used"perceived" post-editing effort labels as quality scores for scoring and ranking in four languages pairs. These labels vary within [1-3], where:

- 1 = perfect translation
- 2 = near miss translation (sentences with 2-3 errors that are easy to fix)
- 3 = very low quality sentence.

The training and test sets for each language pair in Task 1.1 contain 3-4 translations of the same source sentences. The language pairs are German-English (DE-EN) with 150 source sentences for test and 350 source sentences for training, English-German (EN-DE) with 150 source sentences for test and 350 source sentences for training, English-Spanish (EN-ES) with 150 source sentences for test and 954 source sentences for training, and Spanish-English (ES-EN) with 150 source sentences for test and 350 source sentences for training. The translations for each language pair include a human translation and translations produced by a statistical MT (SMT) system, a rule-based MT (RBMT) system, and a hybrid system (for the EN-DE and EN-ES language pairs only).

By inspecting the source side of the training set, we noticed that the translations were ordered per systems, since the source file had sentences repeated in batches. For example, the EN-ES language pair had 954 English sentences and 3,816 Spanish sentences. In the source file, the English sentences were repeated in batches of 954 sentences. Based on that, we assumed that in the target file each set of 954 translations in sequence corresponded to a given MT system (or human).

For each system (human translation is considered as a system, since we do not know the order of the translations), we calculate the consensual information considering the other 2-3 systems available as pseudo-references.

The quality scores for Task 1.2 and Task 1.3 were computed as HTER (Human Translation Error Rate (Snover et al., 2006)) and post-editing time, respectively, for both scoring and ranking.

The datasets were a mixture of test sets from the WMT13 and WMT12 translation shared tasks for the EN-ES language pair only. In this case, the consensual information was extracted by using systems submitted to the WMT translation shared tasks of both years. Therefore, for each source sentence in the WMT12/13 data, all translations produced by the participating MT systems of that year were used as pseudo-references. The *uedin* system outputs for both WMT13 and WMT12 were not considered, since the datasets in Tasks 1.2 and 1.3 were created from translations generated by this system.[2]

The Asyia Toolkit[3] (Giménez and Màrquez, 2010) was used to extract the automatic metrics considered as features. BLEU, TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin and Och, 2004) are used in all task variants. For Tasks 1.2 and 1.3 we also use metrics based on syntactic similarities from shallow and dependency parser information (metrics SPOc(*) and DPmHWCM_c1, respectively, in Asyia). BLEU is a precision-oriented metric that compares n-grams (n=1-4 in our case) from reference documents against n-grams of the MT output, measuring how close the output of a system is to one or more references. TER (Translation Error Rate) measures the minimum number of edits required to transform the MT output into the closest reference document. METEOR (Metric for Evaluation of Translation with Explicit ORdering) scores MT outputs by aligning them with given references. This alignment can be done by exact, stem, synonym and paraphrases matching (here, exact matching was used). ROUGE is a recall-oriented metric that measures similarity between sentences by considering the longest common n-gram statistics between a translation sentence and the corresponding reference sentence. SPOc(*) measures the lexical overlap according to the chunk types of the syntactic realisation. The '*' means that an average of all chunk types is computed. DPmHWCM_c1 is based on the matching of head-word chains. We considered the match of grammatical categories of only one head-word.

These consensual features are combined with the 17 QuEst baseline features provided by the shared task organisers.

[2]WMT14 QE shared task organisers, personal communication, March 2014.

[3]http://asiya.lsi.upc.edu/

3 Experiments and Results

The results reported herein are the official shared task results, that is, they were computed using the true scores of the test set made available by the organisers after our submission.

For training the QE models, we used Support Vector Machines (SVM) regression algorithm with a radial basis function (RBF) kernel with the hyperparameters optimised via grid search. The scikit-learn algorithm available in the QuEst Framework[4] (Specia et al., 2013) was used for that.

We compared the results obtained against using only the QuEst baseline (**BL**) features, which is the same system used as the official baseline for the shared task. For the scoring variant we also compare our results against a baseline that "predicts" the average of the true scores of the training set as scores for each sentence of the test set (**Mean** – each sentence has the same predicted score).

For all language pairs in Task 1.1, Table 1 shows the average results for the scoring variant using MAE (Mean Absolute Error) as evaluation metric, while Table 2 shows the results for the ranking variant using DeltaAvg.

The results for scoring improved over the baselines with the use of consensual information for language pairs DE-EN and EN-ES. For EN-DE and ES-EN the consensual features achieved similar results to BL. The best result for consensual information features was achieved with EN-ES (0.03 of MAE difference from BL).

For the ranking variant, the consensual information improved the results for all language pairs. The largest improvement from consensual-based features was achieved for ES-EN, with a difference of 0.11 from the baseline. It is worth mentioning that for ES-EN our system achieved the best ranking result in Task 1.1.

Since the results varied for different languages pairs, we further inspected them for each language pair. First, we looked at the true scores distribution and realised that the first batch of translations for each language pair was probably the human reference since the percentage of 1s – the best quality score – was much higher for this system (see Figure 1 for EN-DE as an example). By using this human translation as a reference for the other MT systems, we computed BLEU for each sentence

[4]http://www.quest.dcs.shef.ac.uk/

	DE-EN	EN-DE	EN-ES	ES-EN
Mean	0.67	0.68	0.46	0.58
BL	0.65	0.64	0.52	0.57
BL+Consensus	**0.63**	0.64	**0.49**	0.57

Table 1: Scoring results for Task 1.1 in terms of MAE

	DE-EN	EN-DE	EN-ES	ES-EN
BL	0.21	0.23	0.14	0.12
BL+Consensus	**0.28**	**0.26**	**0.21**	**0.23**

Table 2: Ranking results for Task 1.1 in terms of DeltaAvg

and averaged these values. The results are shown in Table 3.

For DE-EN, EN-DE and EN-ES, the various systems appeared to be less dissimilar in terms of BLEU, when compared to ES-EN. For ES-EN, the difference between the two MT systems was higher than for other language pairs (0.12 for the test set and 0.11 for the training set). Moreover, for DE-EN, EN-DE and EN-ES, the difference between the averaged BLEU score of the training set and the average BLEU score of the test set is very small (smaller than 0.01). For ES-EN, however, the difference between the scores for the training and test sets was also higher (0.04 for System1 and 0.03 for System2). This can be one reason why the consensual features did not show improvements for this language pair. Since the systems are considerably different and also there is a considerable difference between training and test sets, the data can be too noisy to be used as pseudo-references.

For EN-DE, the reasons for the bad performance of consensual features are not clear. This language pair showed the worst average quality scores for all systems. Reasons for this can include characteristics of German language, such as compound words which are not well treated in MT, and complex grammar. One hypothesis is that these low BLEU scores (as Table 3 shows) introduce noise instead of useful information for QE. Another difference that appeared only in EN-DE was the distributions of the scores across the different systems. As Figure 1 shows, System1 has a distribution considerably different from the other two systems. For the other language pairs, the distributions across different systems were more uniform. This difference can be another factor influencing the results for this language pair.

Table 4 shows the results for scoring (MAE) and Table 5 shows the results for ranking (DeltaAvg) for Tasks 1.2 and 1.3.

	Task 1.2	Task 1.3
Mean	16.93	23.34
BL	15.23	21.49
BL+Consensus	**13.61**	**21.48**

Table 4: Scoring results of Tasks 1.2 and 1.3 in terms of MAE

	Task 1.2	Task 1.3
BL	5.08	14.71
BL+Consensus	**7.93**	**14.98**

Table 5: Ranking results of Tasks 1.2 and 1.3 in terms of DeltaAvg

For Tasks 1.2 and 1.3 the use of consensual information only slightly improved the baseline results for scoring. For the ranking variant, BL+Consensus achieved better results, but only significantly so for Task 1.2. Therefore, consensual information seems useful to rank sentences according to predicted HTER, its contribution to predicting actual HTER is not noticeable. For post-editing time as quality labels, the improvement achieved with the use of consensual information was marginal.

4 Conclusions

The use of consensual information of MT systems can be useful to improve state-of-the-art results for QE. For some scenarios, it is possible to acquire several translations for a given source segment, but with no additional information on the quality or type of MT systems used to produce them. Therefore, these translations could not be used as pseudo-references in the same way as in (Soricut and Echihabi, 2010).

	DE-EN		EN-DE			EN-ES			ES-EN	
	Sys1	Sys2	Sys1	Sys2	Sys3	Sys1	Sys2	Sys3	Sys1	Sys2
Average BLEU (test)	0.31	0.25	0.20	0.19	0.21	0.36	0.29	0.32	**0.44**	**0.32**
Average BLEU (training)	0.31	0.26	0.21	0.18	0.22	0.35	0.29	0.31	**0.40**	**0.29**

Table 3: Average BLEU of systems in Task 1.1

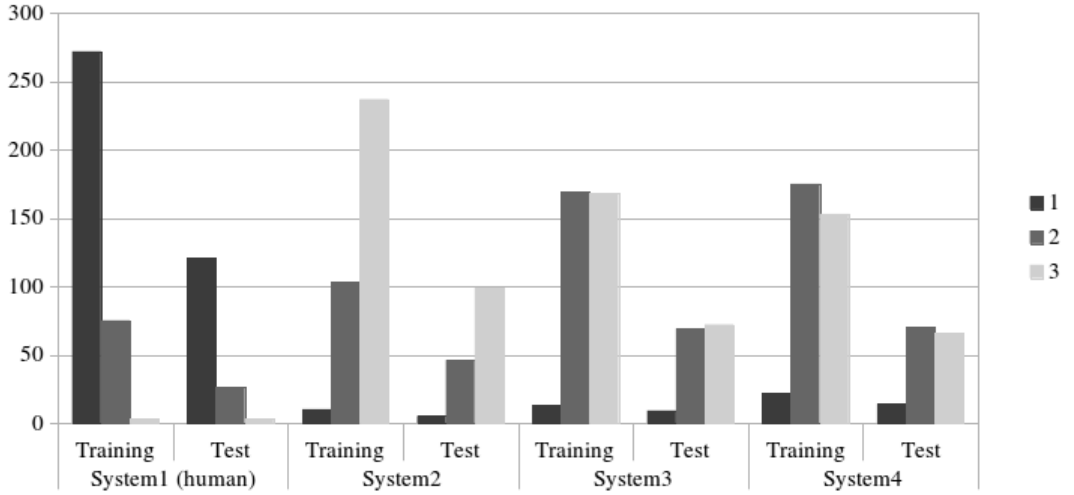

Figure 1: Distribution of true quality scores for the EN-DE language pair

The use of several references with the hypothesis that they share consensual information has been shown useful in some settings, particularly in Task 1.1. In others, the results were inconclusive. In particular, the approach does not seem appropriate for scenarios where the MT systems are considerably different (as shown in Table 3). In those cases, better ways to exploit consensual information need to be investigated further.

Acknowledgements: This work was supported by the EXPERT (EU Marie Curie ITN No. 317471) project.

References

Joshua S. Albrecht and Rebecca Hwa. 2008. The role of pseudo references in mt evaluation. In *Proceedings of WMT 2008*, pages 187–190, Columbus, Ohio, USA.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Chin-Yew Lin and Franz J. Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of ACL 2004*, Barcelona, Spain.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, USA.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the XIV MT Summit*, pages 167–174, Nice, France.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*, pages 223–231.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Transla-

tions via Ranking. In *Proceedings of the ACL 2010*, pages 612–621, Uppsala, Sweden.

Radu Soricut and Sushant Narsale. 2012. Combining Quality Prediction and System Selection for Improved Automatic Translation Output. In *Proceedings of WMT 2012*, Montreal, Canada.

Radu Soricut, Nguyen Bach, and Ziyuan Wang. 2012. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of WMT 2012*, Montreal, Canada.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. Quest - a translation quality estimation framework. In *Proceedings of WMT 2013: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.