UAlacant word-level and phrase-level machine translation quality estimation systems at WMT 2016

Miquel Esplà-Gomis Felipe Sánchez-Martínez Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper describes the Universitat d'Alacant submissions (labeled as UAlacant) to the machine translation quality estimation (MTQE) shared task at WMT 2016, where we have participated in the word-level and phrase-level MTQE subtasks. Our systems use external sources of bilingual information as a *black box* to spot sub-segment correspondences between the source segment and the translation hypothesis. For our submissions, two sources of bilingual information have been used: machine translation (Lucy LT KWIK Translator and Google Translate) and the bilingual concordancer Reverso Context. Building upon the word-level approach implemented for WMT 2015, a method for phrase-based MTQE is proposed which builds on the probabilities obtained for word-level MTQE. For each sub-task we have submitted two systems: one using the features produced exclusively based on online sources of bilingual information, and one combining them with the baseline features provided by the organisers of the task.

1 Introduction

Machine translation quality estimation (MTQE) (Blatz et al., 2004; Specia et al., 2010; Specia and Soricut, 2013) has aroused the interest of both the scientific community and translation companies on account of its noticeable advantages: it can be used to help professional translators in post-editing, to estimate the translation productivity for different translation technologies, or even for budgeting translation projects. In this context, the WMT 2016 MTQE shared task becomes one of the best scenarios in which different approaches to MTQE can be evaluated and compared for different granularities: segment-level (sub-task 1), phrase-level and word-level (sub-task 2), and document-level (sub-task 3).

For the second consecutive year, the submissions of the UAlacant team tackle the word-level MTQE sub-task, but this year they also cover phrase-level MTQE. This year, the shared task featured a dataset obtained by translating segments in English into German using MT, for which it is needed to identify which words and phrases are inadequately translated. In the case of words, this means detecting which words need to be deleted or replaced, while in the case of phrases this means detecting which phrases contain words translated inadequately, but also if there are missing words, or the order of the words in the phrase is not correct. The systems participating in the task are required to apply the labels BAD and OK, either to words or phrases. In this paper we describe the approach behind the submissions of the Universitat d'Alacant team to these sub-tasks. For our word-level submissions we have applied the approach proposed by Esplà-Gomis et al. (2015), where we used black-box bilingual on-line resources. The new task tackles MTQE for translating English into German. For this task we have combined two on-line-available MT systems,[1] Lucy LT KWIK Translator[2] and Google Translate,[3] and the bilingual concordancer Reverso

[1] In the original approach by Esplà-Gomis et al. (2015) Apertium was one of these MT systems, but this year it was replaced since it does not provide a translation system for the languages of the current year's task.

[2] http://www.lucysoftware.com/english/machine-translation/kwik-translator

[3] http://translate.google.com

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 782–786,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

Context[4] to spot sub-segment correspondences between a sentence S in the source language (SL) and a given translation hypothesis T in the target language (TL). As described by Esplà-Gomis et al. (2015), a collection of features is obtained from these correspondences and then used by a binary classifier to determine the final word-level MTQE labels. We have repeated the approach proposed in WMT 2015 for word-level sub-tasks, and have proposed a new one for phrase-level MTQE that builds upon the system trained for word-level MTQE.

The rest of the paper is organised as follows. Section 2 describes the approach used to produce our submissions. Section 3 describes the experimental setting and the results obtained. The paper ends with some concluding remarks.

2 Sources of bilingual information for machine translation quality estimation at the word and phrase levels

The method used to produce the word-level MTQE submissions is the same than that used by the UAlacant team in the last edition of the shared task of MTQE at WMT 2015 (Esplà-Gomis et al., 2015), which uses binary classification based on a collection of information. As in the previous edition of the shared task, we have used online sources of bilingual information to identify sub-segment alignments between the original SL segment S and a given translation hypothesis T in the TL. These sub-segment alignments are identified by: (i) splitting segments S and T in all possible overlapping sub-segments up to a given length L; (ii) using the sources of bilingual information to translate each sub-segment into the other language, i.e. SL sub-segments into TL, and vice versa; and (iii) attempting to match the translated sub-segments either in T or S.

The rest of the section briefly describes the features used for building the submissions both for word-level and phrase-level sub-tasks in the MTQE shared task of WMT 2016.

2.1 Word-level machine translation quality estimation

A complete description of the features used for word-level MTQE can be found in Section 2 of the paper by Esplà-Gomis et al. (2015). We provide here a general description of the type of features used. Esplà-Gomis et al. (2015) describe two types of features: positive and negative ones, i.e. features that would indicate that the current translation is OK, and features that would indicate that it is BAD.

Positive features use those sub-segment pairs (σ, τ) obtained by means of the external sources of bilingual information such that σ matches the source segment S and τ matches the translation hypothesis T. These features provide positive evidence for words in T matching τ. An additional positive feature is defined, which measures the confidence of the sub-segment pairs by using the translation frequency in those sources of bilingual information capable of providing several translation alternatives, such as bilingual concordancers or probabilistic lexicons.

On the other hand, negative features are built from those sub-segment pairs (σ, τ) for which σ fully matches S, but τ matches T only partially. These sub-segment pairs provide negative evidence for those words in T that do not match τ.

2.2 Phrase-level machine translation quality estimation

While the word-level MTQE task has been going on during the last three editions of WMT, this is the first time that this shared task tackles phrase-level MTQE. This problem, as proposed by the organisers of the task, may miss some kinds of errors that are plausible in a phrase, such as missing words (insertions). According to the instructions provided, the organisers describe the problem as follows: *"if a phrase has at least one 'BAD' word, all its labels are replaced with 'BAD'"*; in other words, the problem of phrase-level MTQE just extends the errors found in a given word to the words happening in the same phrase, but does not add new problems related to the new granularity.

The approach proposed for this task builds on the word-level MTQE method described in Section 2.1. In the case of phrase-level MTQE, a binary classifier is also used to classify a phrase either as OK or BAD. This classifier uses the probability of belonging to the class BAD of every word in a phrase as a feature, which is provided by the classifier trained for the task at the word-level. These features are combined with two more binary features, which are aimed at capturing the information provided by the external sources of bilingual information at the level of phrases. Basically, these features take value `true` when the phrase of the translation

[4] `http://context.reverso.net/translation/`

hypothesis being evaluated is confirmed by one or more sources of bilingual information, i.e. if the TL phrase exactly corresponds to a sub-segment in the SL segment. Having two different features allows to capture this information for each translation direction, i.e. if the TL phrase is the result of translating a phrase in the SL, or if the translation of the TL phrase appears as a sub-segment in the SL segment.

Given that phrases have variable lengths (from 1 to 7 words in the data set provided by the organisation), we decided to train specific classifiers for each phrase length using as many features as words in the phrase (plus the two features at the phrase level described above). Alternatively, it would have been possible to experiment with an approach able to deal with sparse features.

3 Submissions to the WMT 2016 shared task on MTQE

This section describes the details of the systems submitted to the MTQE shared task at WMT 2016. This year, the task consisted in estimating the quality of a collection of segments in German that had been obtained through machine translation from English. The organisers provided three datasets:

- *training set*: a collection of 12,000 segments in English (S) and their corresponding machine translations in German (T); for every word/phrase in T, a label was provided: BAD for the words/phrases to be post-edited, and OK for those to be kept unedited;
- *development set*: 1,000 pairs of segments (S, T) with the corresponding MTQE labels, which can be used to optimise the binary classifier trained by using the training set;
- *test set*: 2,000 pairs of segments (S, T) for which the MTQE labels have to be estimated with the binary classifier built on the training and the development sets.

The same data set was used both for word-level and phrase-level MTQE sub-tasks, with the only difference that, for the latter, the limits of the phrases which make up the full translated segments T were provided. In addition, for every sub-task, a collection of baseline features was provided for each word or phrase in T, respectively, in the different datasets. For word-level quality estimation, this collection consists of 22 baseline features, such as the number of occurrences of the word, or part-of-speech information.[5] For phrase-level quality estimation, this collection consists of 72 baseline features, such as the phrase length or its perplexity.[6]

Using these data, four systems have been submitted to the shared task on MTQE at WMT 2016: two for word-level MTQE and two more for phrase-level MTQE. All the systems are based on the binary classifier described bellow in Section 3.1, but using different collections of features. Of the two systems submitted to each sub-task: one was built using only the features described in Section 2, and the other combined them with the baseline features provided by the organisation. Section 3.2 describes the results obtained with each of these approaches by using the following metrics:

- The precision P^c, i.e. the fraction of instances correctly labelled among all the instances labelled as c, where c is the class assigned (either OK or BAD);
- The recall R^c, i.e. the fraction of instances correctly labelled as c among all the instances that should have been labelled as c;
- The F_1^c score, which is defined as
$$F_1^c = \frac{2P^cR^c}{P^c + R^c};$$
and
- The product of F_1^{OK} and F_1^{BAD} scores, which is the main metric used by the organisers of the task for comparing all the submissions made.

3.1 Binary classifier

A *multilayer perceptron* (Duda et al., 2000, Section 6) was used for classification, as implemented in Weka 3.7 (Hall et al., 2009). Following the approach by Esplà-Gomis et al. (2015), the perceptron was built with a single hidden layer containing the same number of nodes as the number of features; this was the best performing architecture in the preliminary experiments.[7] The training sets

[5] The list of features can be found in the file features_list in the package http://www.quest.dcs.shef.ac.uk/wmt16_files_qe/task2_en-de_test.tar.gz

[6] The list of features can be found in the file features_list in the package http://www.quest.dcs.shef.ac.uk/wmt16_files_qe/task2p_en-de_test.tar.gz

[7] The rest of parameters of the classifiers were also kept as in the approach by Esplà-Gomis et al. (2015).

provided by the organisation were used to train the binary classifiers, both for word and phrase levels, while the development sets were used as validation sets on which the training error was computed, in order to minimise the risk of overfitting. The binary classifiers for the sub-task on phrase-level MTQE was trained to optimise the main comparison metric: $F_1^{\mathrm{BAD}} \cdot F_1^{\mathrm{OK}}$, while the classifier for word-level MTQE was trained to optimise the F_1^{BAD} metric, which was the main comparison metric in WMT 2015.[8]

Given that the binary classifier used for the phrase-level sub-task depends on the output of the binary classifier for word-level MTQE, the training process was incremental, training first the word-level MTQE binary classifiers and then the phrase-level ones. It is worth mentioning that the binary classifiers for phrase-level MTQE use the probabilities provided by the best performing system for word-level MTQE: the one that combines the features obtained from on-line sources of bilingual information with the baseline features. However, the phrase-level baseline features are only used in one of the systems submitted.

3.2 Results

Table 1 shows the results obtained by the systems submitted to the shared task on MTQE, both at the level of words and at the level of phrases. The table also includes the results obtained with a binary classifier trained only on the baseline features (baseline), in order to estimate the contribution of the features described in this work on the performance of the system. Incidentally, and in spite of the changes in languages and machine translation systems, the results obtained for word-level MTQE are very similar to those obtained by Esplà-Gomis et al. (2015) for the translation from English into Spanish.

As can also be seen in Table 1, the classifiers using only the baseline features outperform those using only features based on sources of bilingual information, both at the word level and at the phrase level. The difference between both feature families is specially relevant in the case of the phrase-level MTQE. However, the most interesting results are those obtained when combining both feature families. As a result of this combination, an improvement of 5% in F_1^{BAD} and more than 8% in F_1^{OK} with respect to the baseline is obtained for word-level MTQE. In the case of phrase-based MTQE, this improvement is more unbalanced: 1% for F_1^{BAD}, and more than 10% in F_1^{OK}. Therefore, it is possible to conclude that both the baseline features and those obtained from sources of bilingual information are reasonably independent and, therefore, combining them leads to much more successful systems for the two granularities evaluated.

4 Concluding remarks

In this paper we have described the submissions of the Universitat d'Alacant (called UAlacant) team to the sub-task 2 in the MTQE shared task at WMT 2016, which covers the problems of word-level and phrase-level MTQE. Our submissions used on-line available sources of bilingual information in order to obtain features about the translation hypotheses at different granularities. The approach employed is aimed at being system-independent, since it only uses resources produced by external systems, which makes the addition of new sources of bilingual information straightforward. In fact, one of the sources of bilingual information used in the previous edition of the shared task, Apertium, has been replaced by a new one: Lucy LT. The results obtained confirm the conclusion by Esplà-Gomis et al. (2015) that combining the baseline features with those obtained from external sources of bilingual information provide a noticeable improvement, in this case, not only for word-level MTQE, but also for phrase-level MTQE.

Some future work may be interesting, specially as regards the approach to phrase-level MTQE. As already mentioned, it would be interesting to use binary classifiers that support sparse features, in order to be able to directly train a single binary classifier capable to deal with phrases of any length. This would make it possible to put together all the data available, avoiding splitting it into smaller training sets for different classifiers, and therefore allowing to have larger training data set. On the other hand, it may also be interesting to try to use the features defined for word-level MTQE to train the phrase-level MTQE classifier, instead of defining two levels of classification. The main disadvantage of this approach would be the large amount of features, that would make training more expensive.

[8] This optimisation metric was chosen by mistake, following the implementation by Esplà-Gomis et al. (2015); however, when repeating the experiments with the correct optimisation, it was possible to confirm that the difference between the results of the submission and those obtained with the right optimisation metric was not significant.

Granularity	**System**	P^{BAD}	R^{BAD}	F_1^{BAD}	P^{OK}	R^{OK}	F_1^{OK}	$F_1^{\mathrm{OK}} \times F_1^{\mathrm{BAD}}$
word-level	baseline	29.3%	66.4%	40.6%	88.5%	61.6%	72.6%	29.5%
	SBI	28.9%	68.1%	40.6%	88.7%	59.9%	71.5%	29.0%
	SBI+baseline	35.9%	62.4%	45.6%	89.1%	73.4%	80.5%	36.7%
phrase-level	baseline	33.0%	88.7%	48.2%	83.5%	24.2%	37.5%	18.1%
	SBI	30.6%	80.3%	45.9%	82.2%	38.7%	21.3%	9.8%
	SBI+baseline	35.6%	80.3%	49.3%	82.2%	38.7%	52.6%	26.0%

Table 1: Precision (P), recall (R), and F_1 score obtained for the four systems submitted to the shared task on MTQE at WMT 2016. Two of them are based exclusively on the use of *sources of bilingual information* (SBI, see Section 2), and two more combine these SBI with the baseline features provided by the organisers of the task (SBI+baseline). The table also includes the results obtained when training the same binary classifier exclusively on the baseline features (baseline).

Acknowledgements

Work partially funded by the European Commission through project PIAP-GA-2012-324414 (Abu-MaTran) and by the Spanish government through project TIN2015-69632-R (Effortune). We specially thank Reverso-Softissimo and Prompsit Language Engineering for providing the access to the Reverso Context concordancer, the University Research Program for Google Translate that granted us access to the Google Translate service, and Anna Civil from Lucy Software for providing access to the Lucy LT machine translation system.

References

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 315–321.

R. O. Duda, P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. John Wiley and Sons Inc., 2nd edition.

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, and Mikel Forcada. 2015. UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 309–315, Lisbon, Portugal, September. Association for Computational Linguistics.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10–18.

L. Specia and R. Soricut. 2013. Quality estimation for machine translation: preface. *Machine Translation*, 27(3-4):167–170.

L. Specia, D. Raj, and M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.