The UU Submission to the Machine Translation Quality Estimation Task

Oscar Sagemo, Sara Stymne
Uppsala University
Department of Linguistics and Philology
Oscar.Sagemo1914@student.uu.se, sara.stymne@lingfil.uu.se

Abstract

This paper outlines the UU-SVM system for Task 1 of the WMT16 Shared Task in Quality Estimation. Our system uses Support Vector Machine Regression to investigate the impact of a series of features aiming to convey translation quality. We propose novel features measuring reordering and noun translation errors. We show that we can outperform the baseline when we combine it with a subset of our new features.

1 Introduction

In this paper, we describe Uppsala University's submission to the WMT16 shared task in Quality Estimation (QE). Machine Translation Quality Estimation is the task of assessing the quality of a machine translated unit at runtime, without using reference translations. The different units considered for the 2016 shared task in quality estimation are words, phrases and sentences. We participated in task 1, which focuses on sentence-level QE.

Most modern approaches set the task as a regression problem - attempting to accurately predict a continuous quality label through representing translations with feature vectors. The performance of such approaches rely on determining and extracting features that correlate strongly with the proposed quality label and the impact of a wide variety of features. Different types of systems, including system-dependent (glass-box) or system-independent (black-box), linguistically or statistically motivated features, have been explored (Blatz et al., 2004; Quirk, 2004; Specia et al., 2009). The quality label proposed for the sentence-level task is Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), which sets the focus on predicting the post-editing effort needed to correct the translation.

As no information from the MT system used to translate the data was provided, only black-box features can be considered. Furthermore, since the dataset only consists of one translation direction, English-German, language-specific features can be exploited. Our submission proposes novel features attempting to capture some common noun translation errors from English to German as well as measuring the amount of reordering done by the SMT system. These features are combined with more generic linguistically motivated black-box features that improved the prediction accuracy.

2 Features and resources

In this section we will describe the dataset we used and the baseline system. We also give a detailed description of our suggested features.

2.1 Dataset

The dataset for task 1 spans a total of 15,000 English-German translations from the IT domain. Each entry consists of a source segment, its machine translation, a post-edition of the translation and an edit distance score (HTER) derived from the post-edited version. The dataset was split into 12,000 segments as training data, 1,000 for development and 2,000 for testing. The translations were produced by a single in-house MT system from which no system-dependent information was made available for the sentence-level task. These translations were post-edited by professional translators and the HTER was computed using TER(default settings: tokenised, case insensitive, exact matching only, but with scores capped to 100).

In addition to the dataset, we were provided with a set of resources consisting of a language

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 825–830,
Berlin, Germany, August 11-12, 2016.

model (LM), an ngram-counts list of raw ngram occurrences as well as a lexical translation table.

2.2 Baseline system

In order to establish a common ground for measurement, we were provided with a robust baseline system trained with 17 features[1]. The same baseline system has been used for all previous shared tasks in QE and has proven to be well performing across multiple language pairs and text domains (Bojar et al., 2015). The features quantify the complexity of the source sentence and the fluency of the target sentence, by utilizing corpus frequencies, LM probabilites and token counts. We use these 17 baseline features (b17) as the foundation of our system and measure our performance in relation to the baseline system.

2.3 Proposed features

In addition to the provided resources, further tools were used to extract the features: A modified version of the QuEst++ framework, (Specia et al., 2015) with processors and features added and modified where needed, used to extract the baseline features and a majority of our features. Fast_align (Dyer et al., 2013) was used to generate word alignment files. We used Berkeley Parser (Petrov et al., 2006), trained with the included grammars for English and German, to extract phrase structure-based features. We also used SRILM (Stolcke, 2002) to train a Part-Of-Speech (POS) Language Model over the training dataset as well as to compute all LM-based segment probabilities and perplexities. Lastly, we used TreeTagger (Schmid, 1994) trained with the included models for English and German to obtain all POS-related features.

We aimed to obtain consistent features capturing sources of and results of difficulties for SMT systems by quantifying noun translation errors, reordering measures, grammatical correspondence and structural integrity. The following features were considered and tested for inclusion in the feature set for the submission:

Noun Translation Errors In our previous work on English–German SMT (Stymne et al., 2013), we have noted that the translation of noun compounds is problematic. It is common for English compounds, that are written as separate words, to be rendered as separate words or genitive constructions in German, instead of the idiomatic compound. Compounds tend to be common in technical domains, such as IT.

The language-specific scenario in the task setting allowed us to specifically model these issues. We implemented two features attempting to capture these errors in the direction English-German.

- Ratio of Noun groups between source and target
- Ratio of Genitive constructions between source and target

Due to the fact that split compound nouns is a common translation error for German machine translations, we implemented a feature to look for sequences of nouns in target text. The feature looks for any noun group in both source and target and is computed as the ratio of noun groups, where noun groups are defined as the number of occurrences of sequences of two or more nouns.

Another common compound translation is genitive constructions, which can be over-produced in German. We designed a feature that looks for possible genitive constructions in source and target, and is computed as the ratio of genitive constructions, defined as follows:
German: Any noun or proper noun preceeded by a noun and the genitive article *des/der*.
English: Any noun or proper noun preceeded by a noun and the possesive clitic *'s* or the possessive preposition *of*.

Note that these patterns could also match other constructions since "of" can have other uses and "der" is also used for masculine nominative and feminine dative.

Reordering measures Reordering is problematic for MT in general, and for English–German especially for the placement of verbs, which differ between these languages. We explored three metrics that measure the amount of reordering done by the MT system, to investigate a correlation between SMT reordering and edit operations. All metrics are based on alignments between individual words.

- Crossing score: the number of crossings in alignments between source and target
- Kendall Tau distance between alignments in source and target

[1] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

- Squared Kendall Tau distance between alignments in source and target

Crossing score was suggested by Genzel (2010) for SMT reordering and Tau was suggested by Birch and Osborne (2011) for use in a standard metric with a reference translation. To our knowledge we are the first to use these measures for quality estimation. The features are computed over the crossing link pairs in a word alignment file, where the number of crossing links considers crossings of all lengths and the Squared Kendall Tau Distance (SKTD) is defined as shown in Eq. 1.

$$\text{SKTD} = 1 - \sqrt{\frac{|\text{crossing link pairs}|}{|\text{link pairs}|}} \quad (1)$$

Grammatical correspondence We explored several features quantifying grammatical discrepancy, mainly measured in terms of occurences of syntactic phrases or POS tags in accordance with the work of Felice and Specia (2012).

- Ratio of percentage of verb phrases between source and target
- Ratio of percentage of noun phrases between source and target
- Ratio of percentage of nouns between source and target
- Ratio of percentage of pronouns between source and target
- Ratio of percentage of verbs between source and target
- Ratio of percentage of tokens consisting of alphabetic symbols between source and target

Different means of parameterising the relationship between syntactic and POS constituents were explored, we tested the absolute difference, the ratio of occurences as well as the ratio of percentage. We concluded that the ratio of percentage was the preferred metric.

Structural integrity We also investigated features measuring well-formedness as conveyed by syntactic parse trees in line with Avramidis (2012) as well as POS language models

- Source PCFG average confidence of all possible parses in the parser n-best list
- Target PCFG average confidence of all possible parses in the parser n-best list
- Source PCFG log probability
- Target PCFG log probability
- LM log perplexity of POS of the target
- LM log probability of POS of the target

We experimented with different sizes of n-best lists and found that small sizes (1-3) were preferred due to difficulties in coming up with more parse trees for several of the input sentences.

2.4 Learning

As per the baseline system methodology, we use SVM regression (Chang and Lin, 2011) with a Radial Basis Function (RBF) kernel and a grid search algorithm for parameter optimisation, implemented in QuEst++.

3 Experiments

Initial experiments consisted of concatenating features with the baseline set, in order to sort out the features that had a positive impact on performance and disregard the ones that had a negative impact. As per the QuEst++ framework, performance was measured in terms of Mean Average Error (MAE) and Root Mean Square Error (RMSE) which are defined in Eqs. 2 and 3, where x_i , ... , x_n are the values predicted by the SVM model and y_i , ... , y_n are the values provided by the organisers.

$$\text{MAE} = \frac{1}{n}\sum_{i}^{n}|x_i - y_i| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i}^{n}x_i - y_i^2} \quad (3)$$

Positive Impact A majority of the proposed features proved to have a negative impact on the performance metrics through our experiments, leaving only 5/16 features with a positive impact:

	MAE	RMSE
baseline (b17)	13.826	19.507
b17 + Noun Group Ratio	13.759	19.503
b17 + Source PCFG	13.812	19.515
b17 + Target PCFG	13.819	19.534
b17 + Tau	13.801	19.460
b17 + Verb ratio	13.799	19.604
Combined	13.723	19.552

Table 1: Performance in terms of MAE and RMSE for the indiviual features in the Positive Impact set

Feature combinations	MAE	RMSE
baseline	13.826	19.507
+ Source PCFG	13.812	19.515
+ Target PCFG	13.805	19.560
+ Verb ratio	13.795	19.627
+ Tau	13.757	19.522
+ Noun Group Ratio	13.723	19.552

Table 2: Performance in terms of MAE and RMSE for the combined features resulting in the submitted system

- Noun group ratio
- Kendall Tau distance
- Source PCFG log probability
- Target PCFG log probability
- Ratio of percentage of verbs

We present their individual performance when added to the baseline features in Table 1 and when added in combination in Table 2. All these features have an individual positive impact on MAE, whereas only noun group ratio and Tau perform well on RMSE. Furthermore, the noun group ratio and Kendall Tau Distance showed promising results both individually and in combination with our other new features. The verb ratio feature, however, increased RMSE individually but was included in our final system despite this due to its contribution to MAE when combined, as MAE carries a heavier weight in evaluation. Due to time constraints, we did not investigate the relationship between the RMSE and MAE further.

The performance of the novel features in the noun translation errors and reordering measure groups in Table 3. For the reordering features, that are all different ways of measuring the amount of

	MAE	RMSE
baseline (b17)	13.826	19.507
b17 + Crossings	13.834	19.480
b17 + SKTD	13.836	19.468
b17 + Tau	13.801	19.460
b17 + Noun Group Ratio	13.759	19.503
b17 + Genitive constructions	13.840	19.539

Table 3: Performance in terms of MAE and RMSE for the indiviual features describing noun translation errors and reordering

reordering based on word alignments, we notice that only Tau give a positive impact. Our feature for genitive constructions did not give good results.

The surprisingly small amount of positive features may be a result of a disagreement between the proposed features and the data. The features mainly rely on linguistic analyses while the data, being exclusively from the IT-Domain, is inherently irregular. POS- and syntactic phrase-features appears to be particularly unreliable which may be due to the nature of the domain, where series of constituents of uncommon character are frequent, e.g:

Choose File > Save As , and choose
Photoshop DCS 1.0 or Photoshop DCS
2.0 from the Format menu .

↓

Wählen Sie " Bearbeiten " "
Voreinstellungen " (Windows) bzw. "
Bridge CS4 " > " Voreinstellungen " (
Mac OS) und klicken Sie auf "
Miniaturen . "

This appears to especially affect syntactic parsers trained on out-of-domain PCFGs as phrase comparisons were error prone and the parser often had difficulties generating more than 3 trees per sentence. Nevertheless, the probabilities of the parse trees for both source and target slightly increased the performance of the model.

In order to improve the performance of syntactic and POS-related features, a first step would be to use parsers and taggers trained with or adapted to similar in-domain data, as the IT-domain notably differs from the conventional treebanks and corpora commonly used in the field. Furthermore, we think it would be worthwhile to explore the effect of employing dependency parsers rather than

constituency-based parsers for measuring structural integrity and grammatical correspondence.

The amount of reordering done as measured in this paper can suffice to indicate irregularities in reordering through the learning methods. However, simply relying on counting crossings in 1-1 alignments, could inflict noise. All our measures for reordering only measures the difference in word order in a language independent way. For a specific language pair like English–German it would be useful to be able to measure known word order divergences like verb placement, through more carefully designed and targeted measures. A better solution could be adapt the feature to fit the expected reordering for specific translation directions and to quantify it based on infringements of word-order expectations.

4 Conclusion

We trained regression models using a combination of the baseline features and a series of features intended to convey translation quality. We also proposed novel features modeling noun translation errors and reordering amount. A majority of the proposed features were discarded through our experiments with the development data, yet the final feature set was sufficient to surpass the baseline. Of the final features, the noun group ratio showed particularly promising results, as seen in Table 1.

Results were submitted for both the scoring and ranking subtasks of the sentence-level task. The system was, however, intended and optimized for the scoring task. Therefore the ranks were simply defined as the ascending order of the scores with no separate optimization. When computing our model for the final test set, the training scores were capped to an upper bound of 100 and the predicted scores were capped to a lower bound of 0.

In the future we would like to investigate an expanded set of translation errors as well as adapt the concept of reordering measures as features to expected reordering in specific translation directions.

Acknowledgments

This work forms part of the Swedish strategic research programme eSSENCE.

References

Eleftherios Avramidis. 2012. Quality estimation for machine translation output using linguistic analysis and decoding features. In *Proceedings of the seventh workshop on statistical machine translation*, pages 84–90. Association for Computational Linguistics.

Alexandra Birch and Miles Osborne. 2011. Reordering metrics for mt. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1027–1035. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 315. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–649. Association for Computational Linguistics.

Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103. Association for Computational Linguistics.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd international conference on computational linguistics*, pages 376–384. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Christopher Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on*

Language Resources and Evaluation (LREC), pages 825–828, Lisbon, Portugal.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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 association for machine translation in the Americas*, pages 223–231.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, Spain.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, Denver, Colorado, USA.

Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of compound words for statistical machine translation into compounding languages. *Computational Linguistics*, 39(4):1067–1108.