

The UPC Submission to the WMT 2012 Shared Task on Quality Estimation

Daniele Pighin Meritxell González Lluís Màrquez

Universitat Politècnica de Catalunya, Barcelona

{pighin,mgonzalez,lluism}@lsi.upc.edu

Abstract

In this paper, we describe the UPC system that participated in the WMT 2012 shared task on Quality Estimation for Machine Translation. Based on the empirical evidence that fluency-related features have a very high correlation with post-editing effort, we present a set of features for the assessment of quality estimation for machine translation designed around different kinds of n -gram language models, plus another set of features that model the quality of dependency parses automatically projected from source sentences to translations. We document the results obtained on the shared task dataset, obtained by combining the features that we designed with the baseline features provided by the task organizers.

1 Introduction

Quality Estimation (QE) for Machine Translations (MT) is the task concerned with the prediction of the quality of automatic translations in the absence of reference translations. The WMT 2012 shared task on QE for MT (Callison-Burch et al., 2012) required participants to score and rank a set of automatic English to Spanish translations output by a state-of-the-art phrase based machine translation system. Task organizers provided a training dataset of 1, 832 source sentences, together with reference, automatic and post-edited translations, as well as human quality assessments for the automatic translations. Post-editing effort, i.e., the amount of editing required to produce an accurate translation, was selected as the quality criterion, with assessments ranging from 1

(extremely bad) to 5 (good as it is). The organizers also provided a set of linguistic resources and processors to extract 17 global indicators of translation quality (*baseline features*) that participants could decide to employ for their models. For the evaluation, these features are used to learn a baseline predictors for participants to compare against. Systems participating in the evaluation are scored based on their ability to correctly rank the 422 test translations (using DeltaAvg and Spearman correlation) and/or to predict the human quality assessment for each translation (using Mean Average Error - MAE and Root Mean Squared Error - RMSE).

Our initial approach to the task consisted of several experiments in which we tried to identify common translation errors and correlate them with quality assessments. However, we soon realized that simple regression models estimated on the baseline features resulted in more consistent predictors of translation quality. For this reason, we eventually decided to focus on the design of a set of global indicators of translation quality to be combined with the strong features already computed by the baseline system.

An analysis of the Pearson correlation of the baseline features (Callison-Burch et al., 2012)¹ with human quality assessments shows that the two strongest individual predictors of post-editing effort are the n -gram language model perplexities estimated on source and target sentences. This evidence suggests that a reasonable approach to im-

¹Baseline features are also described in <http://www.statmt.org/wmt12/quality-estimation-task.html>.

Feature	Pearson $ r $	Feature	Pearson $ r $
BL/4	0.3618	DEP/C+/Q4/R	0.0749
BL/5	0.3544	BL/13	0.0741
BL/12	0.2823	DEP/C-/Q1/W	0.0726
BL/14	0.2675	DEP/C+/Q4/W	0.0718
BL/2	0.2667	DEP/C+/Q34/R	0.0687
BL/1	0.2620	BL/3	0.0623
BL/8	0.2575	DEP/C+/Q34/W	0.0573
BL/6	0.2143	SEQ/sys-ref/W	0.0495
DEP/C-/S	0.2072	SEQ/sys/W	0.0492
BL/10	0.2033	SEQ/ref-sys/W	0.0390
DEP/C-/Q12/S	0.1858	BL/7	0.0351
BL/17	0.1824	SEQ/sys/SStop	0.0312
BL/16	0.1725	SEQ/sys/RStop	0.0301
DEP/C-/W	0.1584	SEQ/sys-ref/SStop	0.0291
DEP/C-/R	0.1559	SEQ/sys-ref/RStop	0.0289
DEP/C-/Q12/R	0.1447	DEP/Coverage/S	0.0286
DEP/Coverage/W	0.1419	SEQ/ref-sys/S	0.0232
DEP/C-/Q1/S	0.1413	SEQ/ref-sys/R	0.0205
BL/15	0.1368	SEQ/ref-sys/RStop	0.0187
DEP/C+/Q4/S	0.1257	SEQ/sys-ref/R	0.0184
DEP/Coverage/R	0.1239	SEQ/sys/R	0.0177
SEQ/ref-sys/PStop	0.1181	SEQ/ref-sys/Chains	0.0125
SEQ/sys/PStop	0.1173	SEQ/ref-sys/SStop	0.0104
SEQ/sys-ref/PStop	0.1170	SEQ/sys/S	0.0053
DEP/C-/Q12/W	0.1159	SEQ/sys-ref/S	0.0051
DEP/C-/Q1/R	0.1113	SEQ/sys/Chains	0.0032
DEP/C+/Q34/S	0.0933	SEQ/sys-ref/Chains	0.0014
BL/9	0.0889	BL/11	0.0001

Table 1: Pearson correlation (in absolute value) of the baseline (BL) features and the extended feature set (SEQ and DEP) with the quality assessments.

prove the accuracy of the baseline would be to concentrate on the estimation of other n -gram language models, possibly working at different levels of linguistic analysis and combining information coming from the source and the target sentence. On top of that, we add another class of features that capture the quality of grammatical dependencies projected from source to target via automatic alignments, as they could provide clues about translation quality that may not be captured by sequential models.

The novel features that we incorporate are described in full detail in the next section; in Section 3 we describe the experimental setup and the resources that we employ, while in Section 4 we present the results of the evaluation; finally, in Section 5 we draw our conclusions.

2 Extended features set

We extend the set of 17 baseline features with 35 new features:

SEQ: 21 features based on n -gram language models estimated on reference and automatic translations, combining lexical elements of the target sentence and linguistic annotations (POS) automatically projected from the source;

DEP: 18 features that estimate a language model on dependency parse trees automatically projected from source to target via unsupervised alignments.

All the related models are estimated on a corpus of 150K newswire sentences collected from the training/development corpora of previous WMT editions (Callison-Burch et al., 2007; Callison-Burch et al., 2011). We selected this resource because we prefer to estimate the models only on in-domain data. The models for SEQ features are computed based on reference translations (*ref*) and automatic translations generated by the same Moses (Koehn et al., 2007) configuration used by the organizers of this QE task. As features, we encode the perplexity of observed sequences with respect to the two models, or the ratio of these values. For DEP features, we estimate a model that explicitly captures the difference between reference and automatic translations for the same sentence.

2.1 Sequential features (SEQ)

The simplest sequential models that we estimate are 3-gram language models² on the following sequences:

W: (Word), the sequence of words as they appear in the target sentence;

R: (Root), the sequence of the roots of the words in the target;

S: (Suffix) the sequence of the suffixes of the words in the target;

As features, for each automatic translation we encode:

- The perplexity of the corresponding sequence according to automatic (*sys*) translations: for

²We also considered using longer histories, i.e., 5-grams, but since we could not observe any noticeable difference we finally selected the least over-fitting alternative.

example, $SEQ/sys/R$ and $SEQ/sys/W$ are the root-sequence and word-sequence perplexities estimated on the corpus of automatic translations;

- The ratio between the perplexities according to the two sets of translations: for example, $SEQ/ref-sys/S$ is the ratio between the perplexity of suffix-sequences on reference and automatic translations, and $SEQ/sys-ref/S$ is its inverse.³

We also estimate 3-gram language models on three variants of a sequence in which non-stop words (i.e., all words belonging to an open class) are replaced with either:

RStop: the root of the word;

SStop: the suffix of the word;

PStop: the POS of the aligned source word(s).

This last model (PStop) is the only one that requires source/target pairs in order to be estimated. If the target word is aligned to more than one word, we use the ordered concatenation of the source words POS tags; if the word cannot be aligned, we replace it with the placeholder “*”, e.g.: “*el NN de * VBZ JJ en muchos NNS .*”. Also in this case, different features encode the perplexity with respect to automatic translations (e.g., $SEQ/sys/PStop$) or to the ratio between automatic and reference translations (e.g., $SEQ/ref-sys/RStop$).

Finally, a last class of sequences (**Chains**) collapses adjacent stop words into a single token. Content-words or isolated stop-words are not included in the sequence, e.g.: “*mediante_la de_los de_la y_de_las y_la a_los*”. Again, we consider the same set of variants, e.g. $SEQ/sys/Chains$ or $SEQ/sys-ref/Chains$.

Since there are 7 sequence types and 3 combinations (sys , $sys-ref$, $ref-sys$) we end up with 21 new features.

³Features extracted solely from reference translations have been considered, but they were dropped during development since we could not observe a noticeable effect on prediction quality.

2.2 Dependency features (DEP)

These features are based on the assumption that by observing how dependency parses are projected from source to target we can gather clues concerning translation quality that cannot be captured by sequential models. The features encode the extent to which the edges of the projected dependency tree are observed in reference-quality translations.

The model for DEP features is estimated on the same set of 150K English sentences and the corresponding reference and automatic translations, based on the following algorithm:

1. Initialize two maps M^+ and M^- to store edge counts;
2. Then, for each source sentence s : parse s with a dependency parser;
3. Align the words of s with the reference and the automatic translations r and a ;
4. For each dependency relation $\langle d, s_h, s_m \rangle$ observed in the source, where d is the relation type and s_h and s_m are the head and modifier words, respectively:
 - (a) Identify the aligned head/modifier words in r and a , i.e., $\langle r_h, r_m \rangle$ and $\langle a_h, a_m \rangle$;
 - (b) If $r_h = a_h$ and $r_m = a_m$, then increment $M^+_{\langle d, a_h, a_m \rangle}$ by one, otherwise increment $M^-_{\langle d, a_h, a_m \rangle}$.

In other terms, M^+ keeps track of how many times a projected dependency is the same in the automatic and in the reference translation, while M^- accounts for the cases in which the two projections differ.

Let T be the set of dependency relations projected on an automatic translation. In the feature space we represent:

Coverage: The ratio of dependency edges found in M^- or M^+ over the total number of projected edges, i.e.

$$\text{Coverage}(T) = \frac{\sum_{D \in T} M_D^+ + M_D^-}{|T|} ;$$

C⁺: The quantity $C^+ = \frac{1}{|T|} \sum_{D \in T} \frac{M_D^+}{M_D^+ - M_D^-}$;

C^- : The quantity $C^- = \frac{1}{|T|} \sum_{D \in T} \frac{M_D^-}{M_D^+ - M_D^-}$.

Intuitively, high values of C^+ mean that most projected dependencies have been observed in reference translations; conversely, high values of C^- suggest that most of the projected dependencies were only observed in automatic translations.

Similarly to SEQ features, also in this case we actually employ three variants of these features: one in which we use word forms (i.e., *DEP/Coverage/W*, *DEP/C⁺/W* and *DEP/C⁻/W*), one in which we look at roots (i.e., *DEP/Coverage/R*, *DEP/C⁺/R* and *DEP/C⁻/R*) and one in which we only consider suffixes (i.e., *DEP/Coverage/S*, *DEP/C⁺/S* and *DEP/C⁻/S*).

Moreover, we also estimate C^+ in the top (Q4) and top two (Q34) fourths of edge scores, and C^- in the bottom (Q1) and bottom two (Q12) fourths. As an example, the feature *DEP/C⁺/Q4/R* encodes the value of C^+ within the top fourth of the ranked list of projected dependencies when only considering word roots, while *DEP/C⁻/W* is the value of C^- on the whole edge set estimated using word forms.

3 Experiment setup

To extract the extended feature set we use an alignment model, a POS tagger and a dependency parser. Concerning the former, we trained an unsupervised model with the Berkeley aligner⁴, an implementation of the symmetric word-alignment model described by Liang et al. (2006). The model is trained on Europarl and newswire data released as part of WMT 2011 (Callison-Burch et al., 2011) training data. For POS tagging and semantic role annotation we use SVMTool⁵ (Jesús Giménez and Lluís Màrquez, 2004) and Swirl⁶ (Surdeanu and Turmo, 2005), respectively, with default configurations. To estimate the SEQ and DEP features we use reference and automatic translations of the newswire section of WMT 2011 training data. The automatic translations are generated by the same configuration generating the data for the quality estimation task. The n -gram models are estimated with the

⁴<http://code.google.com/p/berkeleyaligner>

⁵<http://www.lsi.upc.edu/~nlp/SVMTool/>

⁶<http://www.surdeanu.name/mihai/swirl/>

Feature set	DeltaAvg	MAE
Baseline	0.4664	0.6346
Extended	0.4694	0.6248

Table 2: Comparison of the baseline and extended feature set on development data.

SRILM toolkit⁷, with order equal to 3 and Kneser-Ney (Kneser and Ney, 1995) smoothing.

As a learning framework we resort to Support Vector Regression (SVR) (Smola and Schölkopf, 2004) and learn a linear separator using the SVM-Light optimizer by Joachims (1999)⁸. We represent feature values by means of their z-scores, i.e., the number of standard deviations that separate a value from the average of the feature distribution. We carry out the system development via 5-fold cross evaluation on the 1,832 development sentences for which we have quality assessments.

4 Evaluation

In Table 1 we show the absolute value of the Pearson correlation of the features used in our model, i.e., the 17 baseline features (BL/*), the 21 sequence (SEQ/*) and the 18 dependency (DEP/*) features, with the human quality assessments. The more correlated features are in the top (left) part of the table. At a first glance, we can see that 9 of the 10 features having highest correlation are already encoded by the baseline. We can also observe that DEP features show a higher correlation than SEQ features. This evidence seems to contradict our initial expectations, but it can be easily ascribed to the limited size of the corpus used to estimate the n -gram models (150K sentences). This point is also confirmed by the fact that the three variants of the *PStop model (based on sequences of target stop-words interleaved by POS tags projected from the source sentence and, hence, on a very small vocabulary) are the three sequential models sporting the highest correlation. Alas, the lack of lexical anchors makes them less useful as predictors of translation quality than BL/4 and BL/5. Another interesting as-

⁷<http://www-speech.sri.com/projects/srilm>

⁸<http://svmlight.joachims.org/>

System	DeltaAvg	MAE
Baseline	0.55	0.69
Official Evaluation	0.22	0.84
Amended Evaluation	0.51	0.71

Table 3: Official and amended evaluation on test data of the extended feature sets.

pect is that DEP/C⁻ features show higher correlation than DEP/C⁺. This is an expected behaviour, as being indicators of possible errors they are intended to have discriminative power with respect to the human assessments. Finally, we can see that more than 50% of the included features, including five baseline features, have negligible (less than 0.1) correlation with the assessments. Even though these features may not have predictive power per se, their combination may be useful to learn more accurate models of quality.⁹

Table 2 shows a comparison of the baseline features against the extended feature set as the average DeltaAvg score and Mean Absolute Error (MAE) on the 10 most accurate development configurations. In both cases, the extended feature set results in slightly more accurate models, even though the improvement is hardly significant.

Table 3 shows the results of the official evaluation. Our submission to the final evaluation (*Official*) was plagued by a bug that affected the values of all the baseline features on the test set. As a consequence, the official performance of the model is extremely poor. The row labeled *Amended* shows the results that we obtained after correcting the problem. As we can see, on both tasks the baseline outperforms our model, even though the difference between the two is only marginal. Ranking-wise, our official submission is last on the ranking task and last-but-one on the quality prediction task. In contrast, the amended model shows very similar accuracy to the baseline, as the majority of the systems that took part in the evaluation.

⁹Our experiments on development data were not significantly affected by the presence or removal of low-correlation features. Given the relatively small feature space, we adopted a conservative strategy and included all the features in the final models.

5 Discussion and conclusions

We have described the system with which we participated in the WMT 2012 shared task on quality estimation. The model incorporates all the baseline features, plus two sets of novel features based on: 1) n -gram language models estimated on mixed sequences of target sentence words and linguistic annotations projected from the source sentence by means of automatic alignments; and 2) the likelihood of the projection of dependency relations from source to target.

On development data we found out that the extended feature set granted only a very marginal improvement with respect to the strong feature set of the baseline. In the official evaluation, our submission was plagued by a bug affecting the generation of baseline features for the test set, and as a result we had an incredibly low performance. After fixing the bug, re-evaluating on the test set confirmed that the extended set of features, at least in the current implementation, does not have the potential to significantly improve over the baseline features. On the contrary, the accuracy of the corrected model is slightly lower than the baseline on both the ranking and the quality estimation task.

During system development it was clear that improving significantly over the results of the baseline features would be very difficult. In our experience, this is especially due to the presence among the baseline features of extremely strong predictors of translation quality such as the perplexity of the automatic translation. We could also observe that the parametrization of the learning algorithm had a much stronger impact on the final accuracy than the inclusion/exclusion of specific features from the model.

We believe that the information that we encode, and in particular dependency parses and stop-word sequences, has the potential to be quite relevant for this task. On the other hand, it may be necessary to estimate the models on much larger datasets in order to compensate for their inherent sparsity. Furthermore, more refined methods may be required in order to incorporate the relevant information in a more determinant way.

Acknowledgments

This research has been partially funded by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914).

References

- [Callison-Burch et al.2007] Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation*. ACL, Prague, Czech Republic.
- [Callison-Burch et al.2011] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, July.
- [Callison-Burch et al.2012] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.
- [Jesús Giménez and Lluís Màrquez2004] Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th LREC*.
- [Joachims1999] Thorsten Joachims. 1999. Making large-scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*.
- [Kneser and Ney1995] Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan, May.
- [Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.
- [Liang et al.2006] Percy Liang, Benjamin Taskar, and Dan Klein. 2006. Alignment by agreement. In *HLT-NAACL*.
- [Smola and Schölkopf2004] Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August.
- [Surdeanu and Turmo2005] Mihai Surdeanu and Jordi Turmo. 2005. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 221–224, Ann Arbor, Michigan, June.