PRHLT Submission to the WMT12 Quality Estimation Task

Jesús González Rubio and **Alberto Sanchis** and **Francisco Casacuberta**
D. Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de vera s/n, 46022, Valencia, Spain
{jegonzalez,josanna,fcn}@dsic.upv.es

Abstract

This is a description of the submissions made by the pattern recognition and human language technology group (PRHLT) of the Universitat Politècnica de València to the quality estimation task of the seventh workshop on statistical machine translation (WMT12). We focus on two different issues: how to effectively combine subsequence-level features into sentence-level features, and how to select the most adequate subset of features. Results showed that an adequate selection of a subset of highly discriminative features can improve efficiency and performance of the quality estimation system.

1 Introduction

Quality estimation (QE) (Ueffing et al., 2003; Blatz et al., 2004; Sanchis et al., 2007; Specia and Farzindar, 2010) is a topic of increasing interest in machine translation (MT). It aims at providing a quality indicator for unseen translations at various granularity levels. Different from MT evaluation, QE do not rely on reference translations and is generally addressed using machine learning techniques to predict quality scores.

Our main focus in this article is in the combination of subsequence features into sentence features, and in the selection of a subset of relevant features to improve performance and efficiency. Section 2 describes the features and the learning algorithm used in the experiments. Section 3 describe two different approaches implemented to select the best-performing subset of features. Section 4 displays the results of the experimentation intended to determine the optimal setup to train our final submission. Finally, section 5 summarizes the submission and discusses the results.

2 Features and Learning Algorithm

2.1 Available Sources of Information

The WMT12 QE task is carried out on English–Spanish news texts produced by a phrase-based MT system. As training data we are given 1832 translations manually annotated for quality in terms of post-editing effort (scores in the range $[1, 5]$), together with their source sentences, decoding information, reference translations, and post-edited translations. Additional training data can be used, as deemed appropriate. Any of these information sources can be used to extract the features, however, test data consists only on source sentence, translation, and search information. Thus, features were extracted from the sources of information available in test data only. Additionally, we compute some extra features from the WMT12 translation task (WMT12TT) training data.

2.2 Features

We extracted a total of 475 features classified into sentence-level and subsequence-level features. We considered subsequences of sizes one to four.

Sentence-level features

- Source and target sentence lengths, and ratio.
- Proportion of dead nodes in the search graph.
- Number of source phrases.
- Number and average size of the translation options under consideration during search.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 104–108,
Montréal, Canada, June 7-8, 2012.

- Source and target sentence probability and perplexities computed by language models of order one to five.
- Target sentence probability, probability divided by sentence length, and perplexities computed by language models of order one to five. Language models were trained on the 1000-best translations.
- 1000-best average sentence length, 1000-best vocabulary divided by average length, and 1000-best vocabulary divided by source sentence length.
- Percentage of subsequences (sizes one to four) previously unseen in the source training data.

Subsequence-level features

- Frequency of source subsequences in the WMT12TT data.
- IBM Model-1 confidence score for each word in the translation (Ueffing et al., 2003).
- Subsequence confidence scores computed on 1000-best translations as described in (Ueffing et al., 2003; Sanchis et al., 2007). We use four subsequence correctness criteria (Levensthein position, target position, average position, and any position) and three weighting schemes (translation probability, translation rank, and relative frequencies).
- Subsequence confidence scores computed by a smoothed naïve bayes classifier (Sanchis et al., 2007). We computed a confidence score for each correctness criteria (Levensthein, target, average and any). The smoothed classifier was tuned to improve classification error rate on a separate development set (union of news-test sets for years 2008 to 2011).

2.3 Combination of Subsequence-level Features

Since WMT12 focuses on sentence-level QE, subsequence-level features must be combined to obtain sentence-level indicators. We used two different methods to combine subsequence features:

- Average value of subsequence-level scores, as done in (Blatz et al., 2004).
- Percentage of subsequence scores belonging to each frequency quartile[1], as done in (Specia and Farzindar, 2010).

Thus, each subsequence-level feature was represented as five sentence-level features: one average score plus four quartile percentages.

Both methods aim at summarizing the scores of the subsequences in a translations. The average is a rough indicator that measures the "middle" value of the scores while the percentages of subsequences belonging to each quartile are more fine-grained indicators that try to capture how spread out the subsequence scores are.

2.4 Learning Algorithm

We trained our quality estimation model using an implementation of support vector machines (Vapnik, 1995) for regression. Specifically, we used SVM$^{\text{light}}$ (Joachims, 2002) for regression with a radial basis function kernel with the parameters C, w and γ optimized. The optimization was performed by cross-validation using ten random subsamples of the training set (1648 samples for training and 184 samples for validation).

3 Feature Selection

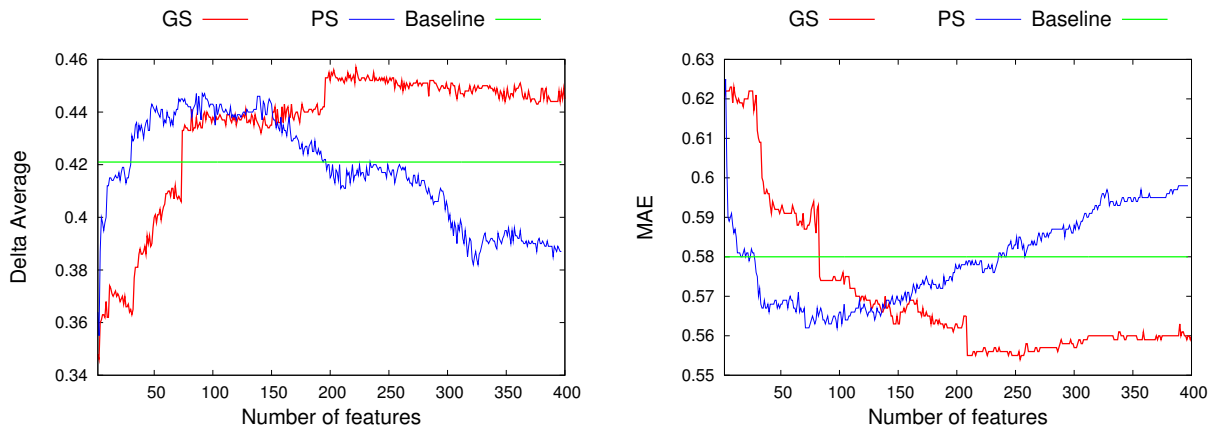
One of the principal challenges that we had to confront is the small size of the training data (only 1832 samples) in comparison with the large number of features, 475. This inadequate amount of training data did not allow for an acceptable training of the regression model which yielded instable systems with poor performance. We also verified that many features were highly correlated and were even redundant sometimes. Since the amount of training data is fixed, we tried to improve the robustness of our regression systems by selecting a subset of relevant features.

We implemented two different feature selection techniques: one based on partial component analysis (PCA), and a greedy selection according to the individual performance of each feature.

3.1 PCA Selection (PS)

Principal component analysis (Pearson, 1901) (PCA) is a mathematical procedure that uses an or-

[1] Quartile values were computed on the WMT12TT data.



(a) Delta average score

(b) Mean Average Error

Figure 1: Delta average score (a) (higher is better) and mean average error (b) (lower is better) as a function of the number of features. Cross-validation results for PCA selection (PS), and greedy selection (GS) methods.

thogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be uncorrelated with the preceding components. Strictly speaking, PCA does not perform a feature selection because the principal components are linear combinations of the individual features.

PCA generates sets of features (the principal components) with almost no correlation. However, it ignores the quality scores to be predicted. Since we want to obtain the best-performing subset of features, there is a mismatch between the selection criterion of PCA and the criterion we are interested in. In other words, although the features generated by PCA contain almost no redundancy, they do not necessarily have to constitute the best-performing subset of features.

3.2 Greedy Performance-driven Selection (GS)

We also implemented a greedy feature selection method which iteratively creates subsets of increasing size with the best-scoring individual features. The score of each feature is given by the performance of a system trained solely on that feature. At a given iteration, we select the K best scoring features and train a regression system with them.

Since we select the features incrementally according to their individual performance, we expect to obtain the subset of features that yield the best performance. However, we do not take into account the correlations that may exist between the different features, thus, the final subset is almost sure to contain a large number of redundant features.

4 Experiments

4.1 Assessment Measures

The organizers propose two variations of the task that will be evaluated separately:

Ranking: Participants are required to submit a ranking of translations. This ranking will used to split the data into n quantiles. The evaluation will be performed in terms of delta average score, the average difference over n between the scores of the top quantiles and the overall score of the corpus. The Spearman correlation will be used as tie-breaking metric.

Scoring: Participants are required to assign a score in the range $[1, 5]$ for each translation. The evaluation will be performed in terms of mean average error (MAE). Root mean squared error (RMSE) will be used as tie-breaking metric.

4.2 Pre-Submission Results

We now describe a number of experiments whose goal is to determine the optimal training setup.

Specifically, we wanted to determine which selection method to use (PCA or greedy) and which features yield a better system. As a preliminary step, we extracted all the features described in section 2. The complete training data consisted on 1832 samples each one with 475 features.

We trained systems using feature sets of increasing size as given by PCA selection (PS) or greedy selection (GS). The parameters of each system were tuned to optimize each of the evaluation measures under consideration. Performance was measured as the average of a ten-fold cross-validation experiment on the training data.

Figure 1 shows the results obtained for the experiments that optimized delta average, and MAE (result optimizing Spearman and RMSE were quite similar). We also display the performance of a system trained on the baseline features. We observed that both selection methods yielded a better performance than the baseline system. PS allowed for a quick improvement in performance as more features are selected, reaching its best results when selecting approximately 80 features. After that, performance rapidly deteriorate. Regarding GS, its improvements in performance were slower in comparison with PS. However, GS finally reached the best scores of the experimentation when selecting $\sim$ 225 features. Specifically, the best performance was reached using the top 222 features for delta average, and using the top 254 features for MAE.

According to these results, our submissions were trained on the best subsets of features as given by the GS method. 222 features were selected according to their delta average score for the ranking task variation, and 254 according to their MAE value for the scoring task variation. Final submissions were trained on the complete training set.

Most of the selected features are sentence-level features calculated from subsequence-based scores. For instance, among the 222 features of the ranking variation of the task, 174 were computed from subsequence scores. Among these 174 features, 129 were calculated from confidence scores computed on 1000-best translations, 29 from confidence scores computed by a smoothed naïve bayes classifier, 11 from the frequencies of the subsequences in the WMT12TT data, and 5 from IBM Model-1 word confidence scores.

Participant ID	Delta average ⇑	MAE ⇓
SDL Language Weaver	0.63	0.61
Uppsala U.	0.58	0.64
LORIA Institute	–	0.68
Trinity College Dublin	0.56	0.68
Baseline	*0.55*	*0.69*
PRHLT	**0.55**	**0.70**
U. Edinburgh	0.54	0.68
Shanghai Jiao Tong U.	0.53	0.69
U. Wolverhampton/Sheffield	0.51	0.69
DFKI	0.46	0.82
Dublin City U.	0.44	0.75
U. Politècnica Catalunya	0.22	0.84

Table 1: Best official evaluation results on each task of the different participating teams. Results for our submissions are displayed in bold. Baseline results in italics.

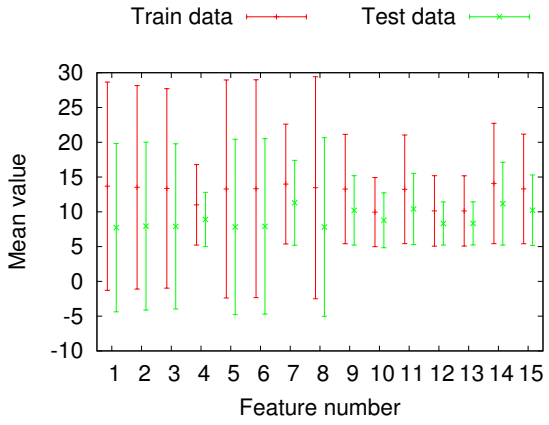


Figure 2: Average value ($\pm$ std. deviation) of the first 15 features used in our final submissions. Feature values follow a similar distribution in the training and test data.

4.3 Official Evaluation Results

After establishing the optimal training setup, we now show the official evaluation results for our submissions. Table 1 shows the performance of the various participants in the ranking (delta average) and scoring (MAE) tasks. Surprisingly our submissions yielded a slightly worse result than the baseline features. However, given the large improvements over the baseline system obtained in the pre-submission experiments, we expected to obtain similar improvements over Baseline in test.

We considered two possible explanations for this counterintuitive result. First, a possibly divergence between the underlying distributions of the training and test data. To investigate this possibility, we stud-

ied the distributions of feature values in the training and test data. Figure 2 displays mean±std. deviation for the first 15 features used in our final submissions (similar results are obtained for all the 222 features). We can observe that feature values in training and test data follow a similar distribution, although test values tend to be slightly lower than training values.

A second plausible explanation is the small amount of training data (only 1832 samples). Limited data favors simpler systems that can train its few free parameters more accurately. This is the case of the Baseline system that was trained using only 11 features, in comparison with the 222 features used in our submissions. Since the training and test data seem to have been generated following the same underlying distribution, we hypothesize that the limited training data is the main explanation for the poor test performance of our submissions.

5 Summary and Discussion

We have presented the submissions of the PRHLT group to the WMT12 QE task. The estimation systems were based on support vector machines for regression. Several features were used to train the systems in order to predict human-annotated post-editing effort scores. Our main focus in this article have been the combination of subsequence features into sentence features, and the selection of a subset of relevant features to improve the submitted systems performance.

Results of the experiments showed that PCA selection was able to obtain better performance when selecting a small number of features while GS yielded the best-performing systems but using much more features. Among the selected features, the larger percentage of them were calculated from subsequence features. These facts indicate that the combination of subsequence features yields sentence-level features with a strong individual performance. However, the high number of features selected by GS indicate that these top-scoring features are highly correlated.

Official evaluation results differ from what we expected; baseline system performs better than our submissions while pre-submission experiments yielded just opposite results. After discarding a possibly discrepancy between training and test data distributions, and given that smaller models such as the baseline system can be trained more accurately with limited data, we concluded that the limited training data is the main explanation for the disparity between our training and test results.

A future line of research could be the study of methods that allow to select sets of uncorrelated features, that unlike PCA, also take into account the individual performance of each feature. Specifically, we plan to study a features selection technique based on partial least squares regression.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 287576. Work also supported by the EC (FEDER/FSE) and the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and iTrans2 (TIN2009-14511) project and by the Generalitat Valenciana under grant ALMPR (Prometeo/2009/01).

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In M. Rollins, editor, *Mental Imagery*. Yale University Press.

Thorsten Joachims. 2002. SVM light.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Alberto Sanchis, Alfons Juan, and Enrique Vidal. 2007. Estimation of confidence measures for machine translation. In *In Procedings of the MT Summit XI*. Springer-Verlag.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. In *AMTA 2010- workshop, Bringing MT to the User: MT Research and the Translation Industry*. The Ninth Conference of the Association for Machine Translation in the Americas, nov.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *In Proceedings of the MT Summit IX*, pages 394–401. Springer-Verlag.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.