DCU-Symantec Submission for the WMT 2012 Quality Estimation Task

Raphael Rubino[†‡], Jennifer Foster[†], Joachim Wagner[†],
Johann Roturier[‡], Rasul Samad Zadeh Kaljahi[†‡], Fred Hollowood[‡]
[†]Dublin City University, [‡]Symantec, Ireland
[†]firstname.lastname@computing.dcu.ie
[‡]firstname_lastname@symantec.com

Abstract

This paper describes the features and the machine learning methods used by Dublin City University (DCU) and SYMANTEC for the WMT 2012 quality estimation task. Two sets of features are proposed: one *constrained*, i.e. respecting the data limitation suggested by the workshop organisers, and one *unconstrained*, i.e. using data or tools trained on data that was not provided by the workshop organisers. In total, more than 300 features were extracted and used to train classifiers in order to predict the translation quality of unseen data. In this paper, we focus on a subset of our feature set that we consider to be relatively novel: features based on a topic model built using the Latent Dirichlet Allocation approach, and features based on source and target language syntax extracted using part-of-speech (POS) taggers and parsers. We evaluate nine feature combinations using four classification-based and four regression-based machine learning techniques.

1 Introduction

For the first time, the WMT organisers this year propose a Quality Estimation (QE) shared task, which is divided into two sub-tasks: scoring and ranking automatic translations. The aim of this workshop is to define useful sets of features and machine learning techniques in order to predict the quality of a machine translation (MT) output T (Spanish) given a source segment S (English). Quality is measured using a 5-point likert scale which is based on post-editing effort, following the scoring scheme:

1. The MT output is incomprehensible
2. About 50-70% of the MT output needs to be edited
3. About 25-50% of the MT output needs to be edited
4. About 10-25% of the MT output needs to be edited
5. The MT output is perfectly clear and intelligible

The final score is a combination of the scores assigned by three evaluators. The use of a 5-point scale makes the scoring task more difficult than a binary classification task where a translation is considered to be either *good* or *bad*. However, if the task is successfully carried out, the score produced is more useful.

Dublin City University and Symantec jointly address the scoring task. For each pair (S, T) of source segment S and machine translation T, we train three classifiers and one classifier combination using the training data provided by the organisers to predict 5-point Likert scores. In this paper, we present the classification results on the test set along with additional results obtained using regression techniques. We evaluate the usefulness of two new sets of features:

1. topic-based features using Latent Dirichlet Allocation (LDA (Blei et al., 2003)),
2. syntax-based features using POS taggers and parsers (Wagner et al., 2009)

The remainder of this paper is organised as follows. In Section 2, we give an overview of all the

features employed in our QE system. Then, in Section 3, we describe the topic and syntax-based features in more detail. Section 4 presents the various classification and regression techniques we explored. Our results are presented and discussed in Section 5. Finally, we summarise and outline our plans in Section 6.

2 Features Overview

In this section, we describe the features used in our QE system. In the first subsection, the features included in our constrained system are presented. In the second subsection, we detail the features included in our unconstrained system. Both of these systems include the 17 baseline features provided for the shared task.

2.1 Constrained System

The constrained system is based only on the data provided by the organisers. We extracted 70 features in total (including the baseline features) and we present them here according to the type of information they capture.

Word and Phrase-Level Features

- **Ratio of source and target segment length**: the number of source words divided by the number of target words
- **Ratio of source and target number of punctuation marks**: the number of source punctuation marks divided by the number of target ones
- **Number of phrases comprising the MT output**: given a phrase-table, we assume that a sentence composed of several phrases indicates uncertainty on the part of the MT system.
- **Average length of source and target phrases**: concatenating short phrases may result in lower fluency compared to the use of longer ones.
- **Ratio of source and target averaged phrase length**
- **Number of source prepositions and conjunctions word**: our assumption here is that segments containing a relatively high number of prepositions and conjunctions may be more complex and difficult to translate.
- **Number of source out-of-vocabulary words**

Language Model Features

All the language models (LMs) used in our work are n-gram LMs with Kneser-Ney smoothing built with the SRI Toolkit (Stolcke, 2002).

- **Backward 2-gram and 3-gram source and target log probabilities**: as proposed by Duchateau et al. (2002)
- **Log probability of target segments on 5-gram MT-output-based LM**: using MOSES (Koehn et al., 2007) trained on the provided parallel corpus, we translated the English side of this corpus into Spanish, assuming that the MT output contains mistakes. This MT output is used to build a LM that models the behavior of the MT system. We assume that for a given MT output, a high n-gram probability (or a low perplexity) of the LM indicates that the MT output contains mistakes.

MT-system Features

- **15 scores provided by *Moses***: phrase-table, language model, reordering model and word penalty (weighted and unweighted)
- **Number of n-bests for each source segment**
- **MT output back-translation**: from Spanish to English using MOSES trained on the provided parallel corpus, scored with TER (Snover et al., 2006), BLEU (Papineni et al., 2002) and the Levenshtein distance (Levenshtein, 1966), based on the source segments as a translation reference

Topic Model Features

- **Probability distribution over topics**: Source and target segment probability distribution over topics for a 10-dimension topic model
- **Cosine distance between source and target topic vectors**

More details about these two features are provided in Section 3.1.

2.2 Unconstrained System

In addition to the features used for the constrained system, a further 238 unconstrained features were included in our *unconstrained* system.

MT System Features

As for our *constrained* system, we use MT output back-translation from Spanish to English, but this time using *Bing Translator*[1] in addition to *Moses*. Each back-translated segment is scored with TER, BLEU and the Levenshtein distance, based on the source segments as a translation reference.

Source Syntax Features

Wagner et al. (2007; 2009) propose a series of features to measure sentence grammaticality. These features rely on a part-of-speech tagger, a probabilistic parser and a precision grammar/parser. We have at our disposal these tools for English and so we apply them to the source data. The features themselves are described in more detail in Section 3.2.

Target Syntax Features

We use a part-of-speech tagger trained on Spanish to extract from the target data the subset of grammaticality features proposed by Wagner et al. (2007; 2009) that are based on POS n-grams. In addition we extract features which reflect the prevalence of particular POS tags in each target segment. These are explained in more detail in Section 3.2 below.

Grammar Checker Features

LANGUAGETOOL (based on (Naber, 2003)) is an open-source grammar and style proofreading tool that finds errors based on pre-defined, language-specific rules. The latest version of the tool can be run in server mode, so individual sentences can be checked and assigned a total number of errors (which may or may not be true positives).[2] This number is used as a feature for each source segment and its corresponding MT output.

3 Topic and Syntax-based Features

In this section, we focus on the set of features that aim to capture *adequacy* using topic modelling and *grammaticality* using POS tagging and syntactic parsing.

[1] http://www.microsofttranslator.com/

[2] The list of English and Spanish rules is available at: http://languagetool.org/languages.

3.1 Topic-based Features

We extract source and target features based on a topic model built using LDA. The main idea in topic modelling is to produce a set of thematic word clusters from a collection of documents. Using the parallel corpus provided for the task, a bilingual corpus is built where each line is composed of a source segment and its translation separated by a space. Each pair of segments is considered as a bilingual document. This corpus is used to train a bilingual topic model after stopwords removal. The resulting model is one set of bilingual topics z containing words w with a probability $p(w_n|z_n, \beta)$ (with n equal to the vocabulary size in the whole parallel corpus). This model can be used to infer the probability distribution of unseen source and target segments over bilingual topics. During the test step, each source segment and its translation are considered individually, as two monolingual documents. This method allows us to compare the source and target topic distributions. We assume that a source segment and its translation share topic similarities.

We propose two ways of using topic-based features for quality estimation: keeping source and target topic vectors as two sets of k features, or computing a vector distance between these two vectors and using one feature only. To measure the proximity of two vectors, we decided to used the *Cosine* distance, as it leads to the best results in terms of classification accuracy. However, we plan to study different metrics in further experiments, like the *Manhattan* or the *Euclidean* distances. Some parameters related to LDA have to be studied more carefully too, such as the number of topics (dimensions in the topic space), the number of words per topic, the Dirichlet hyperparameter α, etc. In our experiments, we built a topic model composed of 10 dimensions using Gibbs sampling with 1000 iterations. We assume that a higher dimensionality can lead to a better repartitioning of the vocabulary over the topics.

Multilingual LDA has been used before in natural language processing, e.g. polylingual topic models (Mimno et al., 2009) or multilingual topic models for unaligned text (Boyd-Graber and Blei, 2009). In the field of machine translation, Tam et al. (2007) propose to adapt a translation and a lan-

guage model to a specific topic using Latent Semantic Analysis (LSA, or Latent Semantic Indexing, LSI (Deerwester et al., 1990)). More recently, some studies were conducted on the use of LDA to adapt SMT systems to specific domains (Gong et al., 2010; Gong et al., 2011) or to extract bilingual lexicon from comparable corpora (Rubino and Linarès, 2011). Extracting features from a topic model is, to the best of our knowledge, the first attempt in machine translation quality estimation.

3.2 Syntax-based Features

Syntactic features have previously been used in MT for confidence estimation and for building automatic evaluation measures. Corston-Oliver et al. (2001) build a classifier using 46 parse tree features to predict whether a sentence is a human translation or MT output. Quirk (2004) uses a single parse tree feature in the quality estimation task with a 4-point scale, namely whether a spanning parse can be found, in addition to LM perplexity and sentence length. Liu and Gildea (2005) measure the syntactic similarity between MT output and reference translation. Albrecht and Hwa (2007) measure the syntactic similarity between MT output and reference translation and between MT output and a large monolingual corpus. Gimenez and Marquez (2007) explore lexical, syntactic and shallow semantic features and focus on measuring the similarity of MT output to reference translation. Owczarzak et al. (2007) use labelled dependencies together with WordNet to avoid penalising valid syntactic and lexical variations in MT evaluation. In what follows, we describe how we make use of syntactic information in the QE task, i.e. evaluating MT output without a reference translation.

Wagner et al. (2007; 2009) use three sources of linguistic information in order to extract features which they use to judge the grammaticality of English sentences:

1. For each POS n-gram (with n ranging from 2 to 7), a feature is extracted which represents the frequency of the least frequent n-gram in the sentence according to some reference corpus. TreeTagger (Schmidt, 1994) is used to produce POS tags.
2. Features provided by a hand-crafted, broad-coverage precision grammar of English (Butt et al., 2002) and a Lexical Functional Grammar parser (Maxwell and Kaplan, 1996). These include whether or not a sentence could be parsed without resorting to robustness measures, the number of analyses found and the parsing time.
3. Features extracted from the output of three probabilistic parsers of English (Charniak and Johnson, 2005), one trained on Wall Street Journal trees (Marcus et al., 1993), one trained on a distorted version of the treebank obtained by automatically creating grammatical error and adjusting the parse trees, and the third trained on the union of the original and distorted versions.

These features were originally designed to distinguish grammatical sentences from ungrammatical ones and were tested on sentences from learner corpora by Wagner et al. (2009) and Wagner (2012). In this work we extract all three sets of features from the source side of our data and the POS-based subset from the target side.[3] We use the publicly available pre-trained TreeTagger models for English and Spanish[4]. The reference corpus used to obtain POS n-gram frequences is the MT translation model training data.[5]

In addition to the POS-based features described in Wagner et al. (2007; 2009), we also extract the following features from the Spanish POS-tagged data: for each POS tag P and target segment T, we extract a feature which is the proportion of words in T that are tagged as P. Two additional features are extracted to represent the proportion of words in T that are assigned more than one tag by the tagger,

[3] Unfortunately, due to time constraints, we were unable to source a suitable probabilistic phrase-structure parser and a precision grammar for Spanish and were thus unable to extract parser-based features for Spanish. We expect that these features would be more useful on the target side than the source side.

[4] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[5] To aid machine learning methods that linearly combine feature values, we add binarised features derived from the raw XLE and POS n-gram features described above, for example we add a feature indicating whether the frequency of the least frequent POS 5-gram is below 10. We base the choice of binary features on (a) decision rules observed in decision trees trained for a binary scoring task and (b) decision rules of simple classifiers (decision trees with just one decision node and 2 leaf nodes) that form a convex hull of optimal classifiers in ROC space.

and the proportion of words in T that are unknown to the tagger.

4 Machine Learning

In this section, we describe the machine learning methods that we experimented with. Our final systems submitted for the shared task are based on classification methods. However, we also performed some experiments with regression methods.

We evaluate the systems on the test set using the official evaluation script and the reference scores. We report the evaluation results as Mean Average Error (MAE) and Root Mean Squared Error (RMSE).

4.1 Classification

In order to apply classification algorithms to the set of features associated with each source and target segment, we rounded the training data scores to the closest integer. We tested several classifiers and empirically chose three algorithms: Support Vector Machine using sequential minimal optimization and RBF kernel (parameters optimized by grid-search) (Platt, 1999), Naive Bayes (John and Langley, 1995) and Random Forest (Breiman, 2001) (the latter two techniques were applied with default parameters). We use the Weka toolkit (Hall et al., 2009) to train the classifiers and predict the scores on the test set. Each method is evaluated individually and then combined by averaging the predicted scores.

4.2 Regression

We applied three different regression techniques: SVM epsilon-SVR with RBF kernel, Linear Regression and M5P (Quinlan, 1992; Wang and Witten, 1997). The two latter algorithms were used with default parameters, whereas SVM parameters (γ, c and ϵ) were optimized by grid-search. We also performed a combination of the three algorithms by averaging the predicted scores. We apply a linear function on the predicted scores S in order to keep them in the correct range (from 1 to 5) as detailed in (1), where S' is the rescaled sentence score, S_{min} is the lowest predicted score and S_{max} is the highest predicted score.

$$S' = 1 + 4 \times \frac{S - S_{min}}{S_{max} - S_{min}} \quad (1)$$

5 Evaluation

Table 1 shows the results obtained by our classification approach on various feature subsets. Note that the two submitted systems used the combined classifier approach with the constrained and unconstrained feature sets. Table 2 shows the results for the same feature combinations, this time using regression rather than classification.

The results of quality estimation using classification methods show that the baseline and the syntax-based features with the classifier combination leads to the best results with an MAE of 0.71 and an RMSE of 0.87. However, these scores are substantially lower than the ones obtained using regression, where the unconstrained set of features with SVM leads to an MAE of 0.62 and an RMSE of 0.78.

It seems that the classification methods are not suitable for this task according to the different sets of features studied. Furthermore, the topic-distance feature is not correlated with the quality scores, according to the regression results. On the other hand, the syntax-based features appear to be the most informative and lead to an MAE of 0.70.

6 Conclusion

We presented in this paper our submission for the WMT12 Quality Estimation shared task. We also presented further experiments using different machine learning techniques and we evaluated the impact of two sets of features - one set which is based on linguistic features extracted using POS tagging and parsing, and a second set which is based on topic modelling. The best results are obtained by our unconstrained system containing all features and using an ϵ-SVR regression method with a Radial Basis Function kernel. This setup leads to a Mean Average Error of 0.62 and a Root Mean Squared Error of 0.78. Unfortunately, we did not submit our best configuration for the shared task.

We plan to continue working on the task of machine translation quality estimation. Our immediate next steps are to continue to investigate the contribution of individual features, to explore feature selection in a more detailed fashion and to apply our best system to other types of data including sentences taken from an online discussion forum.

	SMO		NAIVE BAYES		RANDOM FOREST		Combination	
Features	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
baseline	**0.74**	**0.89**	0.85	1.10	0.84	1.06	0.71	0.88
topic distribution	0.84	1.02	1.09	1.38	0.91	1.15	0.78	0.98
topic distance	0.88	1.11	0.93	1.17	1.04	1.23	0.84	1.04
syntax	0.78	0.97	1.01	1.27	0.83	1.05	0.72	0.90
baseline + topic	0.82	1.01	1.00	1.31	0.84	1.05	0.75	0.95
baseline + syntax	0.76	0.94	1.01	1.25	0.79	0.98	**0.71**	**0.87**
baseline + topic + syntax	0.82	1.04	1.03	1.29	0.79	0.98	0.74	0.93
all constrained	0.99	1.26	1.12	1.46	**0.71**	**0.88**	0.86 ○	1.12 ○
all unconstrained	0.97	1.25	**0.80**	**1.02**	0.79	0.99	0.75 •	0.97 •

Table 1: MAE and RMSE results for different sets of features using three classification methods. The results with ○ and • correspond to the *DCU-SYMC_constrained* and the *DCU-SYMC_unconstrained* systems respectively, submitted for the shared task.

	SVM		LINEAR REG.		M5P		Combination	
Features	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
baseline	0.78	0.93	0.80	0.99	0.73	0.91	0.72	0.88
topic distribution	0.78	0.95	**0.79**	**0.96**	0.80	0.96	0.79	0.95
topic distance	1.38	1.67	1.31	1.62	1.85	2.09	1.00	1.24
syntax	0.70	0.88	0.97	1.22	1.41	1.65	0.76	0.92
baseline + topic	0.78	0.96	1.06	1.31	1.16	1.42	0.88	1.10
baseline + syntax	0.67	0.82	0.90	1.12	2.17	2.38	0.98	1.22
baseline + topic + syntax	0.68	0.84	0.93	1.16	2.12	2.33	0.97	1.21
all constrained	0.83	1.02	0.94	1.18	0.78	0.99	**0.71**	**0.88**
all unconstrained	**0.62**	**0.78**	1.33	1.60	**0.71**	**0.89**	0.73	0.91

Table 2: MAE and RMSE results for different sets of features using three regression methods.

References

J. Albrecht and R. Hwa. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 880–887.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 75–82.

L. Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

M. Butt, H. Dyvik, T. Holloway King, H. Masuichi, and C. Rohrer. 2002. The parallel grammar project. In *Proceedings of the Coling Workshop on Grammar Engineering and Evaluation*.

E. Charniak and M. Johnson. 2005. Course-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 173–180, Ann Arbor.

S. Corston-Oliver, M. Gamon, and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155, Toulouse, France, July.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

J. Duchateau, K. Demuynck, and P. Wambacq. 2002. Confidence scoring based on backward language models. In *Proceedings IEEE international conference on acoustics, speech, and signal processing, ICASSP'2002*, volume 1, pages 221–224.

J. Giménez and L. Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic, June.

Z. Gong, Y. Zhang, and G. Zhou. 2010. Statistical machine translation based on lda. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 286–290.

Z. Gong, G. Zhou, and L. Li. 2011. Improve smt with source-side "topic-document" distributions. In *MT Summit*, pages 496–501.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

G.H. John and P. Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Eleventh conference on uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10-8, pages 707–710.

D. Liu and D. Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

John Maxwell and Ron Kaplan. 1996. An Efficient Parser for LFG. In *Proceedings of LFG-96*, Grenoble.

D. Mimno, H.M. Wallach, J. Naradowsky, D.A. Smith, and A. McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

D. Naber. 2003. A rule-based style and grammar checker. Technical report, Bielefeld University Bielefeld, Germany.

K. Owczarzak, J. van Genabith, and A. Way. 2007. Labelled dependencies in machine translation evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 104–111, Prague, Czech Republic, June.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

J.C. Platt. 1999. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT Press.

R. J. Quinlan. 1992. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore. World Scientific.

C. Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC*, Lisbon, June.

R. Rubino and G. Linarès. 2011. A multi-view approach for term translation spotting. *Computational Linguistics and Intelligent Text Processing*, 6609:29–40.

H. Schmidt. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Natural Language Processing.*

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. 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.

A. Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *InterSpeech*, volume 2, pages 901–904.

Y.C. Tam, I. Lane, and T. Schultz. 2007. Bilingual lsa-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

J. Wagner, J. Foster, and J. van Genabith. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of EMNLP-CoNLL*, pages 112–121, Prague, Czech Republic, June.

J. Wagner, J. Foster, and J. van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490.

J. Wagner. 2012. *Detecting grammatical errors with treebank-induced probabilistic parsers*. Ph.D. thesis, Dublin City University.

Y. Wang and I. H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.