Linguistic Features for Quality Estimation

Mariano Felice
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street
Wolverhampton, WV1 1SB, UK
Mariano.Felice@wlv.ac.uk

Lucia Specia
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
L.Specia@dcs.shef.ac.uk

Abstract

This paper describes a study on the contribution of linguistically-informed features to the task of quality estimation for machine translation at sentence level. A standard regression algorithm is used to build models using a combination of linguistic and non-linguistic features extracted from the input text and its machine translation. Experiments with English-Spanish translations show that linguistic features, although informative on their own, are not yet able to outperform shallower features based on statistics from the input text, its translation and additional corpora. However, further analysis suggests that linguistic information is actually useful but needs to be carefully combined with other features in order to produce better results.

1 Introduction

Estimating the quality of automatic translations is becoming a subject of increasing interest within the Machine Translation (MT) community for a number of reasons, such as helping human translators post-editing MT, warning users about non-reliable translations or combining output from multiple MT systems. Different from most classic approaches for measuring the progress of an MT system or comparing MT systems, which assess quality by contrasting system output to reference translations such as BLEU (Papineni et al., 2002), *Quality Estimation* (QE) is a more challenging task, aimed at MT systems in use, and therefore without access to reference translations.

From the findings of previous work on reference-dependent MT evaluation, it is clear that metrics exploiting linguistic information can achieve significantly better correlation with human judgments on quality, particularly at the level of sentences (Giménez and Màrquez, 2010). Intuitively, this should also apply for quality estimation metrics: while evaluation metrics compare linguistic representations of the system output and reference translations (e.g. matching of n-grams of part-of-speech tags or predicate-argument structures), quality estimation metrics would perform the (more complex) comparison og linguistic representations of the input and translation texts. The hypothesis put forward in this paper is therefore that using linguistic information to somehow contrast the input and translation texts can be beneficial for quality estimation.

We test this hypothesis as part of the WMT-12 shared task on quality estimation. The system submitted to this task (WLV-SHEF) integrates linguistic information to a strong baseline system using only shallow statistics from the input and translation texts, with no explicit information from the MT system that produced the translations. A variant also tests the addition of linguistic information to a larger set of shallow features. The quality estimation problem is modelled as a supervised regression task using Support Vector Machines (SVM), which has been shown to achieve good performance in previous work (Specia, 2011). Linguistic features are computed using a number of auxiliary resources such as parsers and monolingual corpora.

The remainder of this paper is organised as follows. Section 2 gives an overview of previous work

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 96–103,
Montréal, Canada, June 7-8, 2012.

on quality estimation, Section 3 describes the set of linguistic features proposed in this paper, along with general experimental settings, Section 4 presents our evaluation and Section 5 provides conclusions and a brief discussion of future work.

2 Related Work

Reference-free MT quality assessment was initially approached as a *Confidence Estimation* task, strongly biased towards exploiting data from a Statistical MT (SMT) system and the translation process to model the confidence of the system in the produced translation. Blatz et al. (2004) attempted sentence-level assessment using a set of 91 features (from the SMT system input and translation texts) and automatic annotations such as NIST and WER. Experiments on classification and regression using different machine learning techniques produced not very encouraging results. More successful experiments were later run by Quirk (2004) in a similar setting but using a smaller dataset with human quality judgments.

Specia et al. (2009a) used Partial Least Squares regression to jointly address feature selection and model learning using a similar set of features and datasets annotated with both automatic and human scores. Black-box features (i.e. those extracted from the input and translation texts only) were as discriminative as glass-box features (i.e. those from the MT system). Later work using black-box features only focused on finding an appropriate threshold for discriminating 'good' from 'bad' translations for post-editing purposes (Specia et al., 2009b) and investigating more objective ways of obtaining human annotation, such as post-editing time (Specia, 2011).

Recent approaches have started exploiting linguistic information with promising results. Specia et al. (2011), for instance, used part-of-speech (PoS) tagging, chunking, dependency relations and named entities for English-Arabic quality estimation. Hardmeier (2011) explored the use of constituency and dependency trees for English-Swedish/Spanish quality estimation. Focusing on word-error detection through the estimation of WER, Xiong et al. (2010) used PoS tags of neighbouring words and a link grammar parser to detect words that are not connected to the rest of the sentence. Work by Bach et al. (2011) focused on learning patterns of linguistic information (such as sequences of part-of-speech tags) to predict sub-sentence errors. Finally, Pighin and Màrquez (2011) modelled the expected projections of semantic roles from the input text into the translations.

3 Method

Our work focuses on the use of a wide range of linguistic information for representing different aspects of translation quality to complement shallow, system-independent features that have been proved to perform well in previous work.

3.1 Linguistic features

Non-linguistic features, such as sentence length or n-gram statistics, are limited in their scope since they can only account for very shallow aspects of a translation. They convey no notion of meaning, grammar or content and as a result they could be very biased towards describing only superficial aspects. For this reason, we introduce linguistic features that account for richer aspects of translations and are in closer relation to the way humans make their judgments. All of the proposed features, linguistic or not, are MT-system independent.

The proposal of linguistic features was guided by three main aspects of translation: fidelity, fluency and coherence. The number of features that were eventually extracted was inevitably limited by the availability of suitable tools for the language pair at hand, mainly for Spanish. As a result, many of the features that were initially devised could not be implemented (e.g. grammar checking). A total of 70 linguistic features were extracted, as summarised below, where S and T indicate whether they refer to the source/input or translation texts respectively:

- Sentence 3-gram log-probability and perplexity using a language model (LM) of PoS tags [T]
- Number, percentage and ratio of content words (N, V, ADJ) and function words (DET, PRON, PREP, ADV) [S & T]
- Width and depth of constituency and dependency trees for the input and translation texts and their differences [S & T]

- Percentage of nouns, verbs and pronouns in the sentence and their ratios between [S & T]
- Number and difference in deictic elements in [S & T]
- Number and difference in specific types of named entities (person, organisation, location, other) and the total of named entities [S & T]
- Number and difference in noun, verb and prepositional phrases [S & T]
- Number of "dangling" (i.e. unlinked) determiners [T]
- Number of explicit (pronominal, non-pronominal) and implicit (zero pronoun) subjects [T]
- Number of split contractions in Spanish (i.e. *al=a el*, *del=de el*) [T]
- Number and percentage of subject-verb disagreement cases [T]
- Number of unknown words estimated using a spell checker [T]

While many of these features attempt to check for general errors (e.g. subject verb disagreement), others are targeted at usual MT errors (e.g. "dangling" determiners, which are commonly introduced by SMT systems and are not linked to any words) or target language peculiarities (e.g. Spanish contractions, zero subjects). In particular, studying deeper aspects such as different types of subjects can provide a good indication of how natural a translation is in Spanish, which is a pro-drop language. Such a distinction is expected to spot unnatural expressions, such as those caused by unnecessary pronoun repetition.[1]

For subject classification, we identified all VPs and categorised them according to their preceding NPs. Thus, explicit subjects were classified as pronominal (PRON+VP) or non-pronominal (NON-PRON-NP+VP) while implicit subjects only included elided (zero) subjects (i.e. a VP not preceded by an NP).

Subject-verb agreement cases were estimated by rules analysing person, number and gender matches in explicit subject cases, considering also internal NP agreement between determiners, nouns, adjectives and pronouns.[2] Deictics, common coherence indicators (Halliday and Hasan, 1976), were checked against manually compiled lists.[3] Unknown words were estimated using the JMySpell[4] spell checker with the publicly available Spanish (es_ES) OpenOffice[5] dictionary. In order to avoid incorrect estimates, all named entities were filtered out before spell-checking.

TreeTagger (Schmid, 1995) was used for PoS tagging of English texts, while Freeling (Padró et al., 2010) was used for PoS tagging in Spanish and for constituency parsing, dependency parsing and named entity recognition in both languages.

In order to compute n-gram statistics over PoS tags, two language models of general and more detailed morphosyntactic PoS were built using the SRILM toolkit (Stolcke, 2002) on the PoS-tagged AnCora corpus (Taulé et al., 2008).

3.2 Shallow features

In a variant of our system, the linguistic features were complemented by a set of 77 non-linguistic features:

- Number and proportion of unique tokens and numbers in the sentence [S & T]
- Sentence length ratios [S & T]
- Number of non-alphabetical tokens and their ratios [S & T]
- Sentence 3-gram perplexity [S & T]

[1] E.g. (1) *The girl beside me was smiling rather brightly. She thought it was an honor that the exchange student should be seated next to her.* → **La niña a mi lado estaba sonriente bastante bien. Ella pensó que era un honor que el intercambio de estudiantes se encuentra próximo a ella.* (superfluous)
(2) *She is thought to have killed herself through suffocation using a plastic bag.* → **Ella se cree que han matado a ella mediante asfixia utilizando una bolsa de plástico.* (confusing)

[2] E.g. **Algunas de estas personas se convertirá en héroes.* (number mismatch), **Barricadas fueron creados en la calle Cortlandt.* (gender mismatch), **Buena mentirosos están cualificados en lectura.* (internal NP gender and number mismatch).

[3] These included common deictic terms compiled from various sources, such as *hoy*, *allí*, *tú* (Spanish) or *that*, *now* or *there* (English).

[4] http://kenai.com/projects/jmyspell

[5] http://www.openoffice.org/

- Type/Token Ratio variations: corrected TTR (Carroll, 1964), Log TTR (Herdan, 1960), Guiraud Index (Guiraud, 1954), Uber Index (Dugast, 1980) and Jarvis TTR (Jarvis, 2002) [S & T]
- Average token frequency from a monolingual corpus [S]
- Mismatches in opening and closing brackets and quotation marks [S & T]
- Differences in brackets, quotation marks, punctuation marks and numbers [S & T]
- Average number of occurrences of all words within the sentence [T]
- Alignment score (IBM-4) and percentage of different types of word alignments by GIZA++ (from the SMT training alignment model provided)

Our basis for comparison is the set of 17 *baseline features*, which are shallow MT system-independent features provided by the WMT-12 QE shared task organizers.

3.3 Building QE models

We created two main feature sets from the features listed above for the WMT-12 QE shared task:

WLV-SHEF_FS: all features, that is, baseline features, shallow features (Section 3.2) and linguistic features (Section 3.1).

WLV-SHEF_BL: baseline features and linguistic features (Section 3.1).

Additionally, we experimented with other variants of these feature sets using 3-fold cross validation on the training set, such as only linguistic features and only non-linguistic features, but these yielded poorer results and are not reported in this paper.

We address the QE problem as a regression task by building SVM models with an epsilon regressor and a radial basis function kernel using the LibSVM toolkit (Chang and Lin, 2011). Values for the cost, epsilon and gamma parameters were optimized using 5-fold cross validation on the training set.

	MAE ↓	RMSE ↓	Pearson ↑
Baseline	**0.69**	**0.82**	**0.562**
WLV-SHEF_FS	**0.69**	0.85	0.514
WLV-SHEF_BL	0.72	0.86	0.490

Table 1: Scoring performance

The training sets distributed for the shared task comprised $1,832$ English sentences taken from news texts and their Spanish translations produced by an SMT system, Moses (Koehn et al., 2007), which had been trained on a concatenation of Europarl and news-commentaries data (from WMT-10). Translations were accompanied by a quality score derived from an average of three human judgments of post-editing effort using a 1-5 scale.

The models built for each of these two feature sets were evaluated using the official test set of 422 sentences produced in the same fashion as the training set. Two sub-tasks were considered: (i) **scoring** translations using the 1-5 quality scores, and (ii) **ranking** translations from best to worse. While quality scores were directly predicted by our models, sentence rankings were defined by ordering the translations according to their predicted scores in descending order, with no additional criteria to resolve ties other than the natural ordering given by the sorting algorithm.

4 Results and Evaluation

Table 1 shows the official results of our systems in the **scoring** task in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), the metrics used in the shared task, as well as in terms of Pearson correlation.

Results reveal that our models fall slightly below the baseline, although this drop is not statistically significant in any of the cases (paired t-tests for Baseline vs WLV-SHEF_FS and Baseline vs WLV-SHEF_BL yield $p > 0.05$). This may suggest that for this particular dataset the baseline features already cover all relevant aspects of quality on their own, or simply that the representation of the linguistic features is not appropriate for the task. The quality of the resources used to extract the linguistic features may also have been an issue. However, a feature selection method may find a different com-

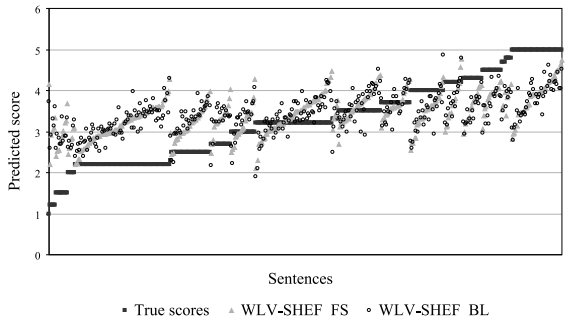

Figure 1: Comparison of true versus predicted scores

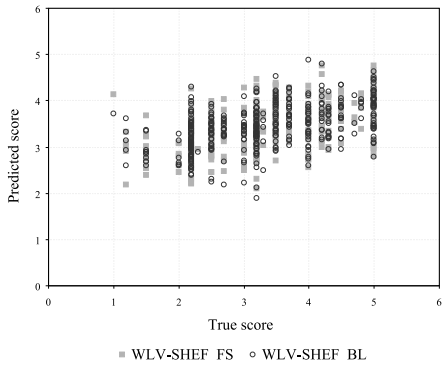

Figure 2: Scatter plot of true versus predicted scores

bination of features that outperforms the baseline, as is later described in this section.

A correlation analysis between our predicted scores and the gold standard (Figure 1) shows some dispersion, especially for the WLV-SHEF_FS set, with lower Pearson coefficients when compared to the baseline. The fluctuation of predicted values for a single score is also very noticeable, spanning more than one score band in some cases. However, if we consider the RMSE achieved by our models, we find that, on average, predictions deviate less than 0.9 absolute points.

A closer look at the score distribution (Figure 2) reveals our models had some difficulty predicting scores in the 1-2 range, possibly affected by the lower proportion of these cases in the training data. In addition, it is interesting to see that the only sentence with a true score of 1 is predicted as a very good translation (with a score greater than 3.5). The reason for this is that the translation has isolated grammatical segments that our features might regard as good but it is actually not faithful to the original.[6] Although the cause for this behaviour can be traced to inaccurate tokenisation, this reveals that our features assess fidelity only superficially and deeper semantically-aware indicators should be explored.

Results for the **ranking** task also fall below the baseline as shown in Table 2, according to the two official metrics: DeltaAvg and Spearman rank correlation coefficient.

	DeltaAvg ↑	**Spearman ↑**
Baseline	**0.55**	**0.58**
WLV-SHEF_FS	0.51	0.52
WLV-SHEF_BL	0.50	0.49

Table 2: Ranking performance

4.1 Further analysis

At first glance, the performance of our models seems to indicate that the integration of linguistic information is not beneficial, since both linguistically-informed feature sets lead to poorer performance as compared to the baseline feature set, which contains only shallow, language-independent features. However, there could be many factors affecting performance so further analysis was necessary to assess their contribution.

Our first analysis focuses on the performance of individual features. To this end, we built and tested models using only one feature at a time and repeated the process afterwards using the full WLV-SHEF_FS set without one feature at a time. In Table 3 we report the 5-best and 5-worst performing features. Although purely statistical features lead the rank, linguistic features also appear among the top five (as indicated by Ⓛ), showing that they can be as good as other shallow features. It is interesting to note that a few features appear as the top performing in both columns (e.g. source bigrams in 4th frequency quartile and target LM probability). These constitute the truly top performing features.

Our second analysis studies the optimal subset of features that would yield the best performance on the test set, from which we could draw further conclusions. Since this analysis requires training and testing models using all the possible partitions of the

[6] *I won't give it away.* → **He ganado ' t darle.*

Rank	One feature	All but one feature
1	*Source bigrams in 4th freq. quartile*	Source average token length
2	Source LM probability	*Source bigrams in 4th freq. quartile*
3	*Target LM probability*	Unknown words in target Ⓛ
4	Number of source bigrams	*Target LM probability*
5	Target PoS LM probability Ⓛ	Difference in constituency tree width Ⓛ
143	Percentage of target S-V agreement Ⓛ	Difference in number of periods
144	Source trigrams in 2nd freq. quartile	Number of source bigrams
145	Target location entities Ⓛ	Target person entities Ⓛ
146	*Source trigrams in 3rd freq. quartile*	Target Corrected TTR
147	Source average translations by inv. freq.	*Source trigrams in 3rd freq. quartile*

Table 3: List of best and worst performing features

full feature set,[7] it is infeasible in practice so we adopted the Sequential Forward Selection method instead (Alpaydin, 2010). Using this method, we start from an empty set and add one feature at a time, keeping in the set only the features that decrease the error until no further improvement is possible. This strategy decreases the number of iterations substantially[8] but it does not guarantee finding a global optimum. Still, a local optimum was acceptable for our purpose. The optimal feature set found by our selection algorithm is shown in Table 4.

Error rates are lower when using this optimal feature set (MAE=0.62 and RMSE=0.76) but the difference is only statistically significant when compared to the baseline with 93% confidence level (paired t-test with $p <= 0.07$). However, this analysis allows us to see how many linguistic features get selected for the optimal feature set.

Out of the total 37 features in the optimal set, 15 are linguistic (40.5%), showing that they are in fact informative when strategically combined with other shallow indicators. This also reveals that feature selection is a key issue for building a quality estimation system that combines linguistic and shallow information. Using a sequential forward selection method, the optimal set is composed of both linguistic and shallow features, reinforcing the idea that they account for different aspects of quality and are not interchangeable but actually complementary.

[7]For 147 features: 2^{147}

[8]For 147 features, worst case is $147 \times (147 + 1)/2 = 10,878$.

5 Conclusions and Future Work

We have explored the use of linguistic information for quality estimation of machine translations. Our approach was not able to outperform a baseline with only shallow features. However, further feature analysis revealed that linguistic features are complementary to shallow features and must be strategically combined in order to be exploited efficiently.

The availability of linguistic tools for processing Spanish is limited, and thus the linguistic features used here only account for a few of the many aspects involved in translation quality. In addition, computing linguistic information is a challenging process for a number of reasons, mainly the fact that translations are often ungrammatical, and thus linguistic processors may return inaccurate results, leading to further errors.

In future work we plan to integrate more global linguistic features such as grammar checkers, along with deeper features such as semantic roles, hybrid n-grams, etc. In addition, we have noticed that representing information for input and translation texts independently seems more appropriate than contrasting input and translation information within the same feature. This representation issue is somehow counter-intuitive and is yet to be investigated.

Acknowledgements

This research was supported by the European Commission, Education & Training, Eramus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

Iter.	Feature
1	Source bigrams in 4th frequency quartile
2	Target PoS LM probability Ⓛ
3	Source average token length
4	Guiraud Index of T
5	Unknown words in T Ⓛ
6	Difference in number of VPs between S and T Ⓛ
7	Diff. in constituency trees width of S and T Ⓛ
8	Non-alphabetical tokens in T
9	Ratio of length between S and T
10	Source trigrams in 4th frequency quartile
11	Number of content words in S Ⓛ
12	Source 3-gram perplexity
13	Ratio of PRON percentages in S and T Ⓛ
14	Number of NPs in T Ⓛ
15	Average number of source token translations with $p > 0.05$ weighted by frequency
16	Source 3-gram LM probability
17	Target simple PoS LM probability Ⓛ
18	Difference in dependency trees depth of S and T Ⓛ
19	Number of NPs in S Ⓛ
20	Number of tokens in S
21	Number of content words in T Ⓛ
22	Source unigrams in 3rd frequency quartile
23	Source unigrams in 1st frequency quartile
24	Source unigrams in 2nd frequency quartile
25	Average number of source token translations with $p > 0.01$ weighted by frequency
26	Ratio of non-alpha tokens in S and T
27	Difference of question marks between S and T normalised by T length
28	Percentage of pron subjects in T Ⓛ
29	Percentage of verbs in T Ⓛ
30	Constituency trees width for S Ⓛ
31	Absolute diff. of question marks between S and T
32	Average num. of source token trans. with $p > 0.2$
33	Diff. of person entities between S and T Ⓛ
34	Diff. of periods between S and T norm. by T length
35	Diff. of semicolons between S and T normalised by T length
36	Source 3-gram perplexity without end-of-sentence markers
37	Absolute difference of periods between S and T

Table 4: An optimal set of features for the test set. The number of iteration indicates the order in which features were selected, giving a rough ranking of features by their performance.

References

Ethem Alpaydin. 2010. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, 2nd edition.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 211–219, Portland, Oregon, USA, June. 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. Final Report of Johns Hopkins 2003 Summer Workshop on Speech and Language Engineering, Johns Hopkins University, Baltimore, Maryland, USA, March.

John Bissell Carroll. 1964. *Language and Thought*. Prentice-Hall, Englewood Cliffs, NJ.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, May.

Daniel Dugast. 1980. *La statistique lexicale*. Slatkine, Genève.

Jesús Giménez and Lluís Màrquez. 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3):209–240.

Pierre Guiraud. 1954. *Les Caractères Statistiques du Vocabulaire*. Presses Universitaires de France, Paris.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Christian Hardmeier. 2011. Improving machine translation quality prediction with syntactic tree kernels. In *Proceedings of the 15th conference of the European Association for Machine Translation (EAMT 2011)*, pages 233–240, Leuven, Belgium.

Gustav Herdan. 1960. *Type-token Mathematics: A Textbook of Mathematical Linguistics*. Mouton & Co., The Hague.

Scott Jarvis. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1):57–84, January.

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.

Llus Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Daniele Pighin and Lluís Màrquez. 2011. Automatic projection of semantic structures: an application to pairwise translation ranking. In *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, Portland, Oregon.

Christopher B. Quirk. 2004. Training a sentence-level machine translation confidence metric. In *Proceedings of the International Conference on Language Resources and Evaluation*, volume 4 of *LREC 2004*, pages 825–828, Lisbon, Portugal.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland, August.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009a. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–35, Barcelona, Spain, May.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009b. Improving the confidence of machine translation quality estimates. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 136–143, Ottawa, Canada, August.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*, pages 19–23, Xiamen, China, September.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.

Andreas Stolcke. 2002. Srilman extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, volume 2, pages 901–904, Denver, USA, November.

Mariona Taulé, M. Antnia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 604–611, Uppsala, Sweden.