Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling

Aaron Li-Feng Han
hanlifengaaron@gmail.com

Yi Lu
mb25435@umac.mo

Derek F. Wong
derekfw@umac.mo

Lidia S. Chao
lidiasc@umac.mo

Liangye He
wutianshui0515@gmail.com

Junwen Xing
mb15470@umac.mo

Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory
Department of Computer and Information Science
University of Macau, Macau S.A.R. China

Abstract

This paper is to introduce our participation in the WMT13 shared tasks on Quality Estimation for machine translation without using reference translations. We submitted the results for Task 1.1 (sentence-level quality estimation), Task 1.2 (system selection) and Task 2 (word-level quality estimation). In Task 1.1, we used an enhanced version of BLEU metric without using reference translations to evaluate the translation quality. In Task 1.2, we utilized a probability model Naïve Bayes (NB) as a classification algorithm with the features borrowed from the traditional evaluation metrics. In Task 2, to take the contextual information into account, we employed a discriminative undirected probabilistic graphical model Conditional random field (CRF), in addition to the NB algorithm. The training experiments on the past WMT corpora showed that the designed methods of this paper yielded promising results especially the statistical models of CRF and NB. The official results show that our CRF model achieved the highest F-score 0.8297 in binary classification of Task 2.

1 Introduction

Due to the fast development of Machine translation, different automatic evaluation methods for the translation quality have been proposed in recent years. One of the categories is the lexical similarity based metric. This kind of metrics includes the edit distance based method, such as WER (Su et al., 1992), Multi-reference WER (Nießen et al., 2000), PER (Tillmann et al., 1997), the works of (Akiba, et al., 2001), (Leusch et al., 2006) and (Wang and Manning, 2012); the precision based method, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and SIA (Liu and Gildea, 2006); recall based method, such as ROUGE (Lin and Hovy 2003); and the combination of precision and recall, such as GTM (Turian et al., 2003), METEOR (Lavie and Agarwal, 2007), BLANC (Lita et al., 2005), AMBER (Chen and Kuhn, 2011), PORT (Chen et al., 2012b), and LEPOR (Han et al., 2012).

Another category is the using of linguistic features. This kind of metrics includes the syntactic similarity, such as the POS information used by TESLA (Dahlmeier et al., 2011), (Liu et al., 2010) and (Han et al., 2013), phrase information used by (Povlsen, et al., 1998) and (Echizen-ya and Araki, 2010), sentence structure used by (Owczarzak et al., 2007); the semantic similarity, such as textual entailment used by (Mirkin et al., 2009) and (Castillo and Estrella, 2012), Synonyms used by METEOR (Lavie and Agarwal, 2007), (Wong and Kit, 2012), (Chan and Ng, 2008); paraphrase used by (Snover et al., 2009).

The traditional evaluation metrics tend to evaluate the hypothesis translation as compared to the reference translations that are usually offered by human efforts. However, in the practice, there is usually no golden reference for the translated documents, especially on the internet works. How to evaluate the quality of automatically translated documents or sentences without using the reference translations becomes a new challenge in front of the NLP researchers.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 365–372,
Sofia, Bulgaria, August 8-9, 2013

ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB		X	.
ADJ	PREP, PREP/DEL	ADV, NEG	CC, CCAD, CCNEG, CQUE, CSUBF, CSUBI, CSUBX	ART	NC, NMEA, NMON, NP, PERCT, UMMX	CARD, CODE, QU	DM, INT, PPC, PPO, PPX, REL	SE	VCLIger, VCLIinf, VCLIfin, VEadj, VEfin, VEger, VEinf, VHadj, VHfin, VHger, VHinf, VLadj,	VLfin, VLger, VLinf, VMadj, VMfin, VMger, VMinf, VSadj, VSfin, VSger, VSinf	ACRNM, ALFP, ALFS, FO, ITJN, ORD, PAL, PDEL, PE, PNC, SYM	BACKSLASH, CM, COLON, DASH, DOTS, FS, LP, QT, RP, SEMICO-LON, SLASH

Table 1: Developed POS mapping for Spanish and universal tagset

2 Related Works

Gamon et al. (2005) perform a research about reference-free SMT evaluation method on sentence level. This work uses both linear and non-linear combinations of language model and SVM classifier to find the badly translated sentences. Albrecht and Hwa (2007) conduct the sentence-level MT evaluation utilizing the regression learning and based on a set of weaker indicators of fluency and adequacy as pseudo references. Specia and Gimenez (2010) use the Confidence Estimation features and a learning mechanism trained on human annotations. They show that the developed models are highly biased by difficulty level of the input segment, therefore they are not appropriate for comparing multiple systems that translate the same input segments. Specia et al. (2010) discussed the issues between the traditional machine translation evaluation and the quality estimation tasks recently proposed. The traditional MT evaluation metrics require reference translations in order to measure a score reflecting some aspects of its quality, e.g. the BLEU and NIST. The quality estimation addresses this problem by evaluating the quality of translations as a prediction task and the features are usually extracted from the source sentences and target (translated) sentences. They also show that the developed methods correlate better with human judgments at segment level as compared to traditional metrics. Popović et al. (2011) perform the MT evaluation using the IBM model one with the information of morphemes, 4-gram POS and lexicon probabilities. Mehdad et al. (2012) use the cross-lingual textual entailment to push semantics into the MT evaluation without using reference translations. This evaluation work mainly focuses on the adequacy estimation. Avramidis (2012) performs an automatic sentence-level ranking of multiple machine translations using the features of verbs, nouns, sentences, subordinate clauses and punctuation occurrences to derive the adequacy information. Other descriptions of the MT Quality Estimation tasks can be gained in the works of (Callison-Burch et al., 2012) and (Felice and Specia, 2012).

3 Tasks Information

This section introduces the different sub-tasks we participated in the Quality Estimation task of WMT 13 and the methods we used.

3.1 Task 1-1 Sentence-level QE

Task 1.1 is to score and rank the post-editing effort of the automatically translated English-Spanish sentences without offering the reference translation.

Firstly, we develop the English and Spanish POS tagset mapping as shown in Table 1. The 75 Spanish POS tags yielded by the Treetagger (Schmid, 1994) are mapped to the 12 universal tags developed in (Petrov et al., 2012). The English POS tags are extracted from the parsed sentences using the Berkeley parser (Petrov et al., 2006).

Secondly, the enhanced version of BLEU (EBLEU) formula is designed with the factors of modified length penalty (MLP), precision, and recall, the h and s representing the lengths of hypothesis (target) sentence and source sentence respectively. We use the harmonic mean of precision and recall, i.e. $H(\alpha R_n, \beta P_n)$. We assign the weight values $\alpha = 1$ and $= 9$, i.e. higher weight value is assigned to precision, which is different with METEOR (the inverse values).

$$EBLEU = 1 - MLP \times exp(\sum w_n log(H(\alpha R_n, \beta P_n))) \quad (1)$$

$$MLP = \begin{cases} e^{1-\frac{s}{h}} & if \ \ h < s \\ e^{1-\frac{h}{s}} & if \ \ h \geq s \end{cases} \quad (2)$$

$$P_n = \frac{\#common\ ngram\ chunk}{\#\ ngram\ chunk\ in\ target\ sentence} \quad (3)$$

$$R_n = \frac{\#common\ ngram\ chunk}{\#\ ngram\ chunk\ in\ source\ sentence} \quad (4)$$

Lastly, the scoring for the post-editing effort of the automatically translated sentences is performed on the extracted POS sequences of the source and target languages. The evaluated performance of EBLEU on WMT 12 corpus is shown in Table 2 using the Mean-Average-Error (MAE), Root-Mean-Squared-Error (RMSE).

	Precision	Recall	MLP	EBLEU
MAE	0.17	0.19	0.25	0.16
RMSE	0.22	0.24	0.30	0.21

Table 2: Performance on the WMT12 corpus

The official evaluation scores of the testing results on WMT 13 corpus are shown in Table 3. The testing results show similar scores as compared to the training scores (the MAE score is around 0.16 and the RMSE score is around 0.22), which shows a stable performance of the developed model EBLEU. However, the performance of EBLEU is not satisfactory currently as shown in the Table 2 and Table 3. This is due to the fact that we only used the POS information as linguistic feature. This could be further improved by the combination of lexical information and other linguistic features such as the sentence structure, phrase similarity, and text entailment.

	MAE	RMSE	DeltaAvg	Spearman Corr
EBLEU	16.97	21.94	2.74	0.11
Baseline SVM	14.81	18.22	8.52	0.46

Table 3: Performance on the WMT13 corpus

3.2 Task 1-2 System Selection

Task 1.2 is the system selection task on EN-ES and DE-EN language pairs. Participants are required to rank up to five alternative translations for the same source sentence produced by multiple translation systems.

Firstly, we describe the two variants of EBLEU method for this task. We score the five alternative translation sentences as compared to the source sentence according to the closeness of their POS sequences. The German POS is also extracted using Berkeley parser (Petrov et al., 2006). The mapping of German POS to universal POS tagset is using the developed one in the work of (Petrov et al., 2012). When we convert the absolute scores into the corresponding rank values, the variant EBLEU-I means that we use five fixed intervals (with the span from 0 to 1) to achieve the alignment as shown in Table 4.

[1,0.4)	[0.4, 0.3)	[0.3, 0.25)	[0.25, 0.2)	[0.2, 0]
5	4	3	2	1

Table 4: Convert absolute scores into ranks

The alignment work from absolute scores to rank values shown in Table 4 is empirically determined. We have made a statistical work on the absolute scores yielded by our metrics, and each of the intervals shown in Table 4 covers the similar number of sentence scores.

On the other hand, in the metric EBLEU-A, "A" means average. The absolute sentence edit scores are converted into the five rank values with the same number (average number). For instance, if there are 1000 sentence scores in total then each rank level (from 1 to 5) will gain 200 scores from the best to the worst.

Secondly, we introduce the NB-LPR model used in this task. NB-LPR means the Naïve Bayes classification algorithm using the features of Length penalty (introduced in previous section), Precision, Recall and Rank values. NB-LPR considers each of its features independently. Let's see the conditional probability that is also known as Bayes' rule. If the $p(x|c)$ is given, then the $p(c|x)$ can be calculated as follows:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad (5)$$

Given a data point identified as $X(x_1, x_2, \dots, x_n)$ and the classifications $C(c_1, c_2, \dots, c_n)$, Bayes' rule can be applied to this statement:

$$p(c_i|x_1, x_2, \dots, x_n) = \frac{p(x_1,x_2,\dots,x_n|c_i)p(c_i)}{p(x_1,x_2,\dots,x_n)} \quad (6)$$

As in many practical applications, parameter estimation for NB-LPR model uses the method of maximum likelihood. For details of Naïve Bayes algorithm, see the works of (Zhang, 2004) and (Harrington, 2012).

Thirdly, the SVM-LPR model means the support vector machine classification algorithm using the features of Length penalty, Precision, Recall and Rank values, i.e. the same features as in NB-LPR. SVM solves the nonlinear classification problem by mapping the data from a low dimensional space to a high dimensional space using the Kernel methods. In the projected high dimensional space, the problem usually becomes a linear one, which is easier to solve. SVM is also called maximum interval classifier because it tries to find the optimized hyper plane that

separates different classes with the largest margin, which is usually a quadratic optimization problem. Let's see the formula below, we should find the points with the smallest margin to the hyper plane and then maximize this margin.

$$arg\,max_{w,b}\left\{min_n(label\cdot(w^Tx+b))\cdot\frac{1}{\|w\|}\right\} \quad (7)$$

where $\boldsymbol{w}$ is normal to the hyper plane, $||\boldsymbol{w}||$ is the Euclidean norm of $\boldsymbol{w}$, and $|b|/||\boldsymbol{w}||$ is the perpendicular distance from the hyper plane to the origin. For details of SVM, see the works of (Cortes and Vapnik, 1995) and (Burges, 1998).

EN-ES					
NB-LPR			SVM-LPR		
MAE	RMSE	Time	MAE	RMSE	Time
.315	.399	.40s	.304	.551	60.67s
DE-EN					
NB-LPR			SVM-LPR		
MAE	RMSE	Time	MAE	RMSE	Time
.318	.401	.79s	.312	.559	111.7s

Table 5: NB-LPR and SVM-LPR training

In the training stage, we used all the officially released data of WMT 09, 10, 11 and 12 for the EN-ES and DE-EN language pairs. We used the WEKA (Hall et al., 2009) data mining software to implement the NB and SVM algorithms. The training scores are shown in Table 5. The NB-LPR performs lower scores than the SVM-LPR but faster than SVM-LPR.

	DE-EN		EN-ES	
Methods	Tau(ties penalized)	|Tau|(ties ignored)	Tau(ties penalized)	|Tau|(ties ignored)
EBLEU-I	-0.38	-0.03	-0.35	0.02
EBLEU-A	N/A	N/A	-0.27	N/A
NB-LPR	-0.49	0.01	N/A	**0.07**
Baseline	-0.12	0.08	-0.23	0.03

Table 6: QE Task 1.2 testing scores

The official testing scores are shown in Table 6. Each task is allowed to submit up to two systems and we submitted the results using the methods of EBLEU and NB-LPR. The performance of NB-LPR on EN-ES language pair shows higher Tau score (0.07) than the baseline system score (0.03) when the ties are ignored. Because of the number limitation of submitted systems for each task, we did not submit the SVM-LPR results. However, the training experiments prove that the SVM-LPR model performs better than the NB-LPR model though SVM-LPR takes more time to run.

3.3 Task 2 Word-level QE

Task 2 is the word-level quality estimation of automatically translated news sentences from English to Spanish without given reference translations. Participants are required to judge each translated word by assigning a two- or multi-class labels. In the binary classification, a good or a bad label should be judged, where "bad" indicates the need for editing the token. In the multi-class classification, the labels include "keep", "delete" and "substitute". In addition to the NB method, in this task, we utilized a discriminative undirected probabilistic graphical model, i.e. Conditional Random Field (CRF).

CRF is early employed by Lefferty (Lefferty et al., 2001) to deal with the labeling problems of sequence data, and is widely used later by other researchers. As the preparation for CRF definition, we assume that X is a variable representing the input sequence, and Y is another variable representing the corresponding labels to be attached to X. The two variables interact as conditional probability $p(Y|X)$ mathematically. Then the definition of CRF: Let a graph model $G=(V,E)$ comprise a set V of vertices or nodes together with a set E of edges or lines and $Y=\{Y_v|v\in V\}$, such that Y is indexed by the vertices of G, then (X,Y) shapes a CRF model. This set meets the following form:

$$P_\theta(Y|X)\propto exp\left(\sum_{e\in E,k}\lambda_k f_k(e,Y|_e,X)+\sum_{v\in V,k}\mu_k g_k(v,Y|_v,X)\right) \quad (8)$$

where X and Y represent the data sequence and label sequence respectively; f_k and g_k are the features to be defined; λ_k and μ_k are the parameters trained from the datasets. We used the tool CRF++[1] to implement the CRF algorithm. The features we used for the NB and CRF are shown in Table 7. We firstly trained the CRF and NB models on the officially released training corpus (produced by Moses and annotated by computing TER with some tweaks). Then we removed the truth labels in the training corpus (we call it pseudo test corpus) and labeled each word using the derived training models. The test results on the pseudo test corpus are shown in Table 8,

[1] https://code.google.com/p/crfpp/

which specifies CRF performs better than NB algorithm.

$U_n, n \in (-4, 3)$	Unigram, from antecedent 4th to subsequent 3rd token
$B_{n-1,n}, n \in (-1, 2)$	Bigram, from antecedent 2nd to subsequent 2nd token
$B_{-1,1}$	Jump bigram, antecedent and subsequent token
$T_{n-1,n,n+1}, n \in (-1, 1)$	Trigram, from antecedent 2nd to subsequent 2nd token

Table 7: Developed features

Binary			
CRF		NB	
Training	Accuracy	Training	Accuracy
Itera=108 Time=2.48s	0.944	Time=0.59s	0.941
Multi-classes			
CRF		NB	
Training	Accuracy	Training	Accuracy
Itera=106 Time=3.67s	0.933	Time=0.55s	0.929

Table 8: Performance on pseudo test corpus

The official testing scores of Task 2 are shown in Table 9. We include also the results of other participants (CNGL and LIG) and their approaches.

	Binary			Multiclass
Methods	Pre	Recall	F1	Acc
CNGL-dMEMM	0.7392	0.9261	0.8222	0.7162
CNGL-MEMM	0.7554	0.8581	0.8035	0.7116
LIG-All	N/A	N/A	N/A	0.7192
LIG-FS	0.7885	0.8644	0.8247	**0.7207**
LIG-BOOSTING	0.7779	0.8843	0.8276	N/A
NB	**0.8181**	0.4937	0.6158	0.5174
CRF	0.7169	**0.9846**	**0.8297**	0.7114

Table 9: QE Task 2 official testing scores

The results show that our method CRF yields a higher recall score than other systems in binary judgments task, and this leads to the highest F1 score (harmonic mean of precision and recall). The recall value reflects the loyalty to the truth data. The augmented feature set designed in this paper allows the CRF to take the contextual information into account, and this contributes much to the recall score. On the other hand, the accuracy score of CRF in multiclass evaluation is lower than LIG-FS method.

4 Conclusions

This paper describes the algorithms and features we used in the WMT 13 Quality Estimation tasks. In the sentence-level QE task (Task 1.1), we develop an enhanced version of BLEU metric, and this shows a potential usage for the traditional evaluation criteria. In the newly proposed system selection task (Task 1.2) and word-level QE task (Task 2), we explore the performances of several statistical models including NB, SVM, and CRF, of which the CRF performs best, the NB performs lower than SVM but much faster than SVM. The official results show that the CRF model yields the highest F-score 0.8297 in binary classification judgment of word-level QE task.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 017/2009/A and RG060/09-10S/CS/FST. The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Akiba, Yasuhiro, Kenji Imamura, and Eiichiro Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of the MT Summit VIII*, Santiago de Compostela, Spain.

Albrecht, Joshua, and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. *ACL*. Vol. 45. No. 1.

Avramidis, Eleftherios. 2012. Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs. In *Proceedings of 24th International Conference on Computational Linguistics* (COLING), pages 115–132, Mumbai, India.

Burges, Christopher J. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. J. *Data Min. Knowl. Discov*. Volume 2 Issue 2, June 1998, 121-167. Kluwer Academic Publishers Hingham, MA, USA.

Callison-Burch, Chris, 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, pages 10–51, Montréal, Canada, June.

Castillo, Julio and Paula Estrella. 2012. Semantic Textual Similarity for MT evaluation, *Proceedings of the 7th Workshop on Statistical Machine Translation* (WMT2012), pages 52–58, Montreál, Canada, June 7-8. Association for Computational Linguistics.

Chan, Yee Seng and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL 2008: HLT*, pages 55–62. Association for Computational Linguistics.

Chen, Boxing and Roland Kuhn. 2011. Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the Sixth Workshop on Statistical Machine translation of the Association for Computational Linguistics(ACL-WMT)*, pages 71-77, Edinburgh, Scotland, UK.

Chen, Boxing, Roland Kuhn and Samuel Larkin. 2012. PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 930–939, Jeju, Republic of Korea, 8-14 July.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-Vector Networks, J. *Machine Learning*, Volume 20, issue 3, pp 273-297. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Dahlmeier, Daniel, Chang Liu, and Hwee Tou Ng. 2011. TESLA at WMT2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics* (ACL-WMT), pages 78-84, Edinburgh, Scotland, UK.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* (HLT '02). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 138-145.

Echizen-ya, Hiroshi and Kenji Araki. 2010. Automatic evaluation method for machine translation using noun-phrase chunking. In *Proceedings of ACL 2010*, pages 108–117. Association for Computational Linguistics.

Gamon, Michael, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. *Proceedings of EAMT*.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11.

Han, Aaron Li-Feng, Derek F. Wong and Lidia S. Chao. 2012. LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. *Proceedings of the 24th International Conference on Computational Linguistics* (COLING 2012: Posters), Mumbai, India.

Han, Aaron Li-Feng, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing and Xiaodong Zeng. 2013. Language-independent Model for Machine Translation Evaluation with Reinforced Factors. *Proceedings of the 14th International Conference of Machine Translation Summit* (MT Summit 2013), Nice, France.

Harrington, Peter. 2012. Classifying with probability theory: naïve bayes. *Machine Learning in Action*, Part 1 Classification. Page 61-82. Publisher: Manning Publications. April.

Lafferty, John, McCallum Andrew, and Pereira C.N. Ferando. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceeding of 18th International Conference on Machine Learning*. 282-289.

Lavie, Alon and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, *Proceedings of the ACL Second Workshop on Statistical Machine Translation*, pages 228-231, Prague, June.

Leusch, Gregor, Nicola Ueffing, and Hermann Ney. 2006. CDer: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL-06), 241-248.

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May 27 - June 1.

Lita, Lucian Vlad, Monica Rogati and Alon Lavie. 2005. BLANC: Learning Evaluation Metrics for MT, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 740–747, Vancouver, October. Association for Computational Linguistics.

Liu, Chang, Daniel Dahlmeier and Hwee Tou Ng. 2010. TESLA: Translation evaluation of sentences

with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Liu, Ding and Daniel Gildea. 2006. Stochastic iterative alignment for machine translation evaluation. Sydney. *ACL06*.

Mariano, Felice and Lucia Specia. 2012. Linguistic Features for Quality Estimation. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 96–103.

Mehdad, Yashar, Matteo Negri, and Marcello Federico. 2012. Match without a referee: evaluating MT adequacy without reference translations. *Proceedings of the Seventh Workshop on Statistical Machine Translation. Association for Computational Linguistics*.

Mirkin, Shachar, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-Language Entailment Modeling for Translating Unknown Terms, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 791–799, Suntec, Singapore, 2-7. ACL and AFNLP.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. A Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC-2000).

Owczarzak, Karolina, Josef van Genabith and Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation, *Proceedings of the ACL Second Workshop on Statistical Machine Translation*, pages 104-111, Prague.

Papineni, Kishore, 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* (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318.

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

Popovic, Maja, David Vilar, Eleftherios Avramidis, Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics* (ACL-WMT), pages 99-103, Edinburgh, Scotland, UK.

Povlsen, Claus, Nancy Underwood, Bradley Music, and Anne Neville. 1998. Evaluating Text-Type Suitability for Machine Translation a Case Study on an English-Danish System. *Proceedings of the First Language Resources and Evaluation Conference, LREC-98*, Volume I. 27-31. Granada, Spain.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Snover, Matthew G., Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. J. *Machine Tranlslation*, 23: 117-127.

Specia, Lucia and Gimenez, J. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. The *Ninth Conference of the Association for Machine Translation in the Americas* (AMTA).

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. Machine Translation Evaluation Versus Quality Estimation. *Machine Translation*, 24:39–50.

Su, Keh-Yih, Wu Ming-Wen and Chang Jing-Shin. 1992. A New Quantitative Quality Measure for Machine Translation Systems. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 433–439, Nantes, France, July.

Tillmann, Christoph, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology* (EUROSPEECH-97).

Turian, Joseph P., Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX*, pages 386–393. International Association for Machine Translation.

Wang, Mengqiu and Christopher D. Manning. 2012. SPEDE: Probabilistic Edit Distance Metrics for MT Evaluation, *WMT2012*, 76-83.

Wong, Billy T. M. and Chunyu Kit. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. *Proceedings of the 2012 Joint Conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning, pages 1060–1068, Jeju Island, Korea, 12–14 July. Association for Computational Linguistics.

Zhang, Harry. 2004. The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, USA. AAAI Press.