All in Strings: a Powerful String-based Automatic MT Evaluation Metric with Multiple Granularities

Junguo Zhu[1], Muyun Yang[1], Bo Wang[2], Sheng Li[1], Tiejun Zhao[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology
{jgzhu; ymy; tjzhao; lish}@mtlab.hit.edu.cn
[2] School of Computer Science and Technology, Tianjin University
bo.wang.1979@gmail.com

Abstract

String-based metrics of automatic machine translation (MT) evaluation are widely applied in MT research. Meanwhile, some linguistic motivated metrics have been suggested to improve the string-based metrics in sentence-level evaluation. In this work, we attempt to change their original calculation units (granularities) of string-based metrics to generate new features. We then propose a powerful string-based automatic MT evaluation metric, combining all the features with various granularities based on SVM rank and regression models. The experimental results show that i) the new features with various granularities can contribute to the automatic evaluation of translation quality; ii) our proposed string-based metrics with multiple granularities based on SVM regression model can achieve higher correlations with human assessments than the state-of-art automatic metrics.

1 Introduction

The automatic machine translation (MT) evaluation has aroused much attention from MT researchers in the recent years, since the automatic MT evaluation metrics can be applied to optimize MT systems in place of the expensive and time-consuming human assessments. The state-of-art strategy to automatic MT evaluation metrics estimates the system output quality according to its similarity to human references. To capture the language variability exhibited by different reference translations, a tendency is to include deeper linguistic information into machine learning based automatic MT evaluation metrics, such as syntactic and semantic information (Amigò et al., 2005; Albrecht and Hwa, 2007; Giménez and Màrquez, 2008). Generally, such efforts may achieve higher correlation with human assessments by including more linguistic features. Nevertheless, the complex and variously presented linguistic features often prevents the wide application of the linguistic motivated metrics.

Essentially, linguistic motivated metrics introduce additional restrictions for accepting the outputs of translations (Amigó et al., 2009). With more linguistic features attributed, the model is actually capturing the sentence similarity in a finer granularity. In this sense, the practical effect of employing various linguistic knowledge is changing the calculation units of the matching in the process of the automatic evaluation.

Similarly, the classical string-based metrics can be changed in their calculation units directly. For example, the calculation granularity in BLEU (Papineni et al., 2002) metric is word: n-grams are extracted on the basis of single word as well as adjacent multiple words. And the calculation granularity in PosBLEU (Popović and Ney, 2009) metric is Pos tag, which correlate well with the human assessments. Therefore, it is straight forward to apply the popular string-based automatic evaluation metrics, such as BLEU, to compute the scores of the systems outputs in the surface or linguis-

Coling 2008: Poster Volume, pages 1533–1540,
Beijing, August 2010

tic tag sequences on various granularities levels.

In this paper, we attempt to change the original calculation units (granularities) of string-based metrics to generate new features. After that, we propose a powerful string-based automatic MT evaluation metric, combining all the features with various granularities based on SVM rank (Joachims, 2002) and regression (Drucker et al., 1996) models. Our analysis indicates that: i) the new features with various granularities can contribute to the automatic evaluation of translation quality; ii) our proposed string-based metrics with multiple granularities based on SVM regression model can achieve higher correlations with human assessments than the state-of-art automatic metrics .

The remainder of this paper is organized as follows: Section 2 reviews the related researches on automatic MT evaluation. Section 3 describes some new calculation granularities of string-based metrics on sentence level. In Section 4, we propose string-based metrics with multiple granularities based on SVM rank and regression models. In Section 5, we present our experimental results on different sets of data. And conclusions are drawn in the Section 6.

2 Related Work on Automatic Machine Translation Evaluation

The research on automatic string-based machine translation (MT) evaluation is targeted at a widely applicable metric of high consistency to the human assessments. WER (Nießen et al., 2000), PER (Tillmann et al., 1997), and TER (Snover et al., 2006) focuses on word error rate of translation output. GTM (Melamed et al., 2003) and the variants of ROUGE (Lin and Och, 2004) concentrate on matched longest common substring and discontinuous substring of translation output according to the human references. BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) are both based on the number of common n-grams between the translation hypothesis and human reference translations of the same sentence. BLEU and NIST are widely adopted in the open MT evaluation campaigns; however, the NIST MT evaluation in 2005 indicates that they can even error in the system level (Le and Przybocki, 2005). Callison-Burch et al. (2006) detailed the deficits of the BLEU and other similar metrics, arguing that the simple surface similarity calculation between the machines translations and the human translations suffers from morphological issues and fails to capture what are important for human assessments.

In order to attack these problems, some metrics have been proposed to include more linguistic information into the process of matching, e.g., Meteor (Banerjee and Lavie, 2005) metric and MaxSim (Chan nad Ng, 2008) metrics, which improve the lexical level by the synonym dictionary or stemming technique. There are also substantial studies focusing on including deeper linguistic information in the metrics (Liu and Gildea, 2005; Owczarzak et al., 2006; Amigó et al., 2006; Mehay and Brew, 2007; Giménez and Màrquez, 2007; Owczarzak et al., 2007; Popovic and Ney, 2007; Giménez and Màrquez, 2008b).

A notable trend improving the string-based metric is to combine various deeper linguistic information via machine learning techniques in the metrics (Amigò et al., 2005; Albrecht and Hwa, 2007; Giménez and Màrquez, 2008). Such efforts are practically amount of introducing additional linguistic restrictions into the automatic evaluation metrics (Amigó et al, 2009), achiving a higher performance at the cost of lower adaptability to other languages owing to the language dependent linguistics features.

Previous work shows that including the new features into the evaluation metrics may benefit to describe nature language accurately. In this sense, the string-based metrics will be improved, if the finer calculation granularities are introduced into the metrics.

Our study analyzes the role of the calculation granularities in the performance of metrics. We find that the new features with various granularities can contribute to the automatic evaluation of translation quality. Also we propose a powerful string based automatic MT evaluation metric with multiple granularities combined by SVM. Finally, we seek a finer feature set of metrics with multiple calculation granularities.

3 The New Calculation Granularities of String-based Metrics on Sentence Level

The string-based metrics of automatic machine translation evaluation on sentence level adopt a common strategy: taking the sentences of the documents as plain strings. Therefore, when changing the calculation granularities of the string-based metrics we can simplify the information of new granularity with plain strings. In this work, five kinds of available calculation granularities are defined: "Lexicon", "Letter", "Pos", "Constitute" and "Dependency".

Lexicon: The calculation granularity is common word in the sentences of the documents, which is popular practice at present.

Letter: Split the granularities of "Lexical" into letters. Each letter is taken as a matching unit.

Pos: The Pos tag of each "Lexicon" is taken as a matching unit in this calculation granularity.

Constitute: Syntactic Constitutes in a tree structure are available through the parser tools. We use Stanford Parser (Klein and Manning, 2003a; Klein and Manning, 2003b) in this work. The Constitute tree is changed into plain string, travelling by BFS (Breadth-first search traversal) [1].

Dependency: Dependency relations in a dependency structure are also available through the parser tools. The dependency structure can also be formed in a tree, and the same processing of being changed into plain string is adopted as "Constitute".

The following serves as an example:

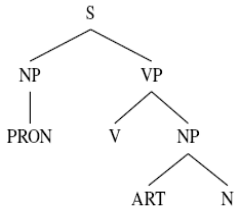
Sentence:

I have a dog

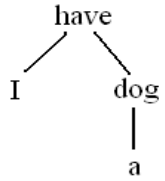
Pos tag:

I/PRON have/V a/ART dog/N

Constitute tree:



Dependency tree:



Then, we can change the sentence into the plain string in multiple calculation granularities as follows:

Lexicon string:

I have a dog

Letter string:

I h a v e a d o g

Pos string:

PRON V ART N

Constitute string:

PRON V ART N NP NP VP S

Dependency string:

a I dog have

The translation hypothesis and human reference translations are both changed into those strings of various calculation granularities. The strings are taken as inputs of the string-based automatic MT evaluation metrics. The outputs of each metric are calculated on different matching units.

4 String-based Metrics with Multiple Granularities Combined by SVM

Introducing machine learning methods to established MT evaluation metric is a popular trend. Our study chooses rank and regression support vector machine (SVM) as the learning model. Features are important for the SVM models.

Plenty of scores can be generated from the proposed metrics. In fact, not all these features are needed. Therefore, feature selection should be a necessary step to find a proper feature set and alleviate the language dependency by using fewer linguistic features.

Feature selection is an NP-Complete problem; therefore, we adopt a greedy selection algorithm called "Best One In" to find a local optimal feature set. Firstly, we select the feature among all the features which best correlates with the human assessments. Secondly, a feature among the rest features is added in to the feature set, if the correlation with the human assessments of the metric using new set is

[1] We also attempt some other traversal algorithms, including preorder, inorder and postorder traversal, the performance are proved to be similar.

the highest among all new metrics and higher than the previous metric in cross training corpus. The cross training corpus is prepared by dividing the training corpus into five parts. Each four parts of the five are for training and the rest one for testing; then, we integrate scores of the five tests as scores of cross training corpus. The five-fold cross training can help to overcome the overfitting. At the end, the feature selection stops, if adding any of the rest features cannot lead to higher correlation with human assessments than the current metric.

5 Experiments

5.1 The Impact of the Calculation Granularities on String-based Metrics

In this section, we use the data from NIST Open MT 2006 evaluation (LDC2008E43), which is described in Table 1. It consists of 249 source sentences that were translated by four human translators as well as 8 MT systems. Each machine translated sentence was evaluated by human judges for their adequacy on a 7-point scale.

	NIST 2002	**NIST 2003**	**NIST Open MT 2006**
LDC corpus	LDC2003 T17	LDC2006 T04	LDC2008 E43
Type	Newswire	Newswire	Newswire
Source	Chinese	Chinese	Arabic
Target	English	English	English
# of sentences	878	919	249
# of systems	3	7	8
# of references	4	4	4
Score	1-5, adequacy & fluency	1-5, adequacy & fluency	1-7 adequacy

Table 1: Description of LDC2006T04, LDC2003T17 and LDC2008E43

To judge the quality of a metric, we compute Spearman rank-correlation coefficient, which is a real number ranging from -1 (indicating perfect negative correlations) to +1 (indicating perfect positive correlations), between the metric's scores and the averaged human assessments on test sentences.

We select 21 features in "lexicon" calculation granularity and 11×4 in the other calculation granularities. We analyze the correlation with human assessments of the metrics in multiple calculation granularities. Table 2 lists the optimal calculation granularity of the multiple metrics on sentence level in the data (LDC2008E43).

Metric	**Granularity**
BLEU-opt	Letter
NIST-opt	Letter
GTM(e=1)	Dependency
TER	Letter
PER	Lexicon
WER	Dependency
ROUGE-opt	Letter

Table 2 The optimal calculation granularity of the multiple metrics

The most remarkable aspect is that not all the best metrics are based on the "lexicon" calculation granularities, such as the "letter" and "dependency". In other words, the granularities-shifted string-based metrics are promising to contribute to the automatic evaluation of translation quality.

5.2 Correlation with Human Assessments of String-based Metrics with Multiple Granularities Based on SVM Frame

We firstly train the SVM rank and regression models on LDC2008E43 using all the features ($21+11 \times 4$ species), without any selection. Secondly, the other two SVM rank and regression models are trained on the same data using the feature set via feature selection, which are described in Table 3. We have four string-based evaluation metrics with multiple granularities on rank and regression SVM frame "Rank_All, Regression_All, Rank_Select and Regression_Select". Then we apply the four metrics to evaluate the sentences of the test data (LDC2006T04 and LDC2003T17). The results of Spearman correlation with human assessments is summarized in Table 3. For comparison, the results from some state-of-art metrics (Papineni et al., 2002; Doddington,

2002; Melamed et al., 2003; Banerjee and Lavie, 2005; Snover et al., 2006; Liu and Gildea, 2005) and two machine learning methods (Albrecht and Hwa, 2007; Ding Liu and Gildea, 2007) are also included in Table 3. Of the two machine learning methods, both trained on the data LDC2006T04. The "Albrecht, 2007" score reported a result of Spearman correlation with human assessments on the data LDC2003T17 using 53 features, while the "Ding Liu, 2007" score reported that under five-fold cross validation on the data LDC2006T04 using 31 features.

	Feature number	**LDC 2003 T17**	**LDC 2006 T04**
Rank_All	65	0.323	0.495
Regression_All	65	**0.345**	0.507
Rank_Select	16	0.338	0.491
Regression_Select	8	0.341	**0.510**
Albrecht, 2007	53	0.309	--
Ding Liu, 2007	31	--	0.369
BLEU-opt[2]	--	0.301	0.453
NIST-opt	--	0.219	0.417
GTM(e=1)	--	0.270	0.375
METEOR[3]	--	0.277	0.463
TER	--	-0.250	-0.302
STM-opt	--	0.205	0.226
HWCM-opt	--	0.304	0.377

Table 3: Comparison of Spearman correlations with human assessments of our proposed metrics and some start-of-art metrics and two machine learning methods
"-opt" stands for the optimum values of the parameters on the metrics

Table 3 shows that the string-based meta-evaluation metrics with multiple granularities based on SVM frame gains the much higher Spearman correlation than other start-of-art metrics on the two test data and, furthermore, our proposed metrics also are higher than the machine learning metrics (Albrecht and Hwa, 2007; Ding Liu and Gildea, 2007).

The underlining is that our proposed metrics are more robust than the aforementioned two machine learning metrics. As shown in Table 1 the heterogeneity between the training and test data in our method is much more significant than that of the other two machine learning based methods.

In addition, the "Regression_Select" metric using only 8 features can achieve a high correlation rate which is close to the metric proposed in "Albrecht, 2007" using 53 features, "Ding Liu, 2007" using 31 features, "Regression_All" and "Rank_All" metrics using 65 features and "Rank_Select" metric using 16 features. What is more, "Regression_Select" metric is better than "Albrecht, 2007", and slightly lower than "Regression_All" on the data LDC2003T17; and better than both "Regression_All" and "Rank_All" metrics on the data LDC2006T04. That confirms that a small cardinal of feature set can also result in a metric having a high correlation with human assessments, since some of the features represent the redundant information in different forms. Eliminating the redundant information is benefit to reduce complexity of the parameter searching and thus improve the metrics performance based on SVM models. Meanwhile, fewer features can relieve the language dependency of the machine learning metrics. At last, our experimental results show that regression models perform better than rank models in the string-based metrics with multiple granularities based on SVM frame, since "Regression_Select" and "Regression_All" achieve higher correlations with human assessments than the others.

5.3 Reliability of Feature Selection

The motivation of feature selection is keeping the validity of the feature set and alleviating the language dependency. We also look forward to the higher Spearman correlation on the test data with a small and proper feature set.

We use SVM-Light (Joachims, 1999) to train our learning models using NIST Open MT 2006 evaluation data (LDC2008E43), and test on the two sets of data, NIST's 2002 and 2003 Chinese MT evaluations. All the data are described in Table 1. To avoid the bias in the distributions of the two judges' assessments in NIST's 2002 and 2003 Chinese MT evaluations, we normalize the scores following (Blatz et al., 2003).

[2] The result is computed by mteval11b.pl.
[3] The result is computed by meteor-v0.7.

We trace the process of the feature selection. The selected feature set of the metric based on SVM rank includes 16 features and that of the metric based on SVM regression includes 8 features. The selected features are listed in Table 4. The values in Table 4 are absolute Spearman correlations with human assessments of each single feature score. The prefixes "C_", "D_", "L_", "P_", and "W_" represent "Constitute", "Dependency", "Letter", "Pos" and "Lexicon" respectively.

Rank	spearman	Regression	spearman
C_PER	.331	C_PER	.331
C_ROUGE-W	.562	C_ROUGE-W	.562
D_NIST9	.479	D_NIST9	.479
D_ROUGE-W	.679	D_ROUGE-L	.667
L_BLEU6	.702	L_BLEU6	.702
L_NIST9	.691	L_NIST9	.691
L_ROUGE-W	.634	L_ROUGE-W	.634
P_PER	.370	P_ROUGE-W	.683
P_ROUGE-W	.616		
W_BLEU1_ind	.551		
W_BLEU2	.659		
W_GTM	.360		
W_METEOR	.693		
W_NIST5	.468		
W_ROUGE1	.642		
W_ROUGE-W	.683		

Table 4: Feature sets of SVM rank and regression

Table 4 shows that 8 features are selected from 65 features in the process of feature selection based on SVM regression while 16 features based on SVM rank. Fewer features based on SVM regression are selected than SVM rank. Only one feature in feature set based on SVM regression does not occur in that based on SVM rank. The reason is that there are more complementary advantages between the common selected features.

Next, we will verify the reliability of our feature selection algorithm. Figure 1 and Figure 2 show the Spearman correlation values between our SVM-based metrics (regression and rank) and the human assessments on both training data (LDC2008E43) and test data (LDC2006T04 and LDC2003T17).

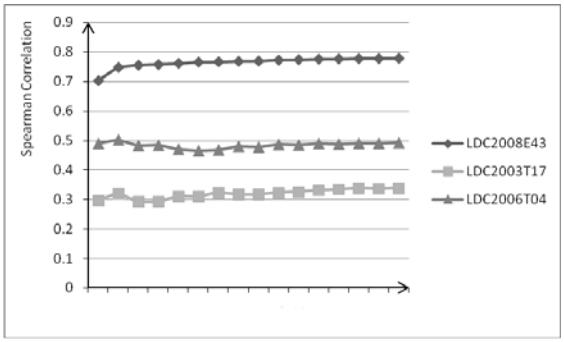


Figure 1: The Spearman correlation values between our **SVM rank** metrics and the human assessments on both training data and test data with the extension of the feature sets

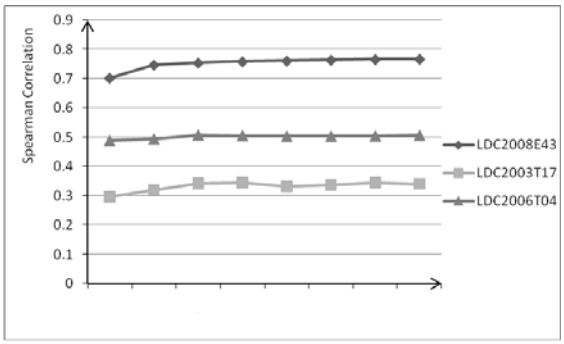


Figure 2: The Spearman correlation values between our **SVM regression** metrics and the human assessments on both training data and test data with the extension of the feature sets

From Figure 1 and Figure 2, with the extension of the feature sets, we can find that the tendency of correlation obtained by each metric based on SVM rank or regression roughly the same on both the training data and test data. Therefore, the two feature sets of SVM rank and regression models are reliable.

6 Conclusion

In this paper we propose an integrated platform for automatic MT evaluation by improving the string based metrics with multiple granularities. Our proposed metrics construct a novel integrated platform for automatic MT evaluation based on multiple features. Our key contribution consists of two parts: i) we suggest a strategy of changing the various complex features into plain string form. According to the strategy, the automatic MT evaluation frame are

much more clarified, and the computation of the similarity is much more simple, since the various linguistic features may express in the uniform strings with multiple calculation granularities. The new features have the same form and are dimensionally homogeneous; therefore, the consistency of the features is enhanced strongly. ii) We integrate the features with machine learning and proposed an effective approach of feature selection. As a result, we can use fewer features but obtain the better performance.

In this framework, on the one hand, string-based metrics with multiple granularities may introduce more potential features into automatic evaluation, with no necessarily of new similarity measuring method, compared with the other metrics. On the other hand, we succeed in finding a finer and small feature set among the combinations of plentiful features, keeping or improving the performance. Finally, we proposed a simple, effective and robust string-based automatic MT evaluation metric with multiple granularities.

Our proposed metrics improve the flexibility and performance of the metrics based on the multiple features; however, it still has some drawbacks: i) some potential features are not yet considered, e.g. the semantic roles; and ii) the loss of information exists in the process of changing linguistic information into plain strings. For example, the dependency label in the calculation granularity "Dependency" is lost when changing information into string form. Though the final results obtain the better performance than the other linguistic metrics, the performance is promising to be further improved if the loss of information can be properly dealt with.

Acknowledgement

This work is supported by Natural Science foundation China (Grant No.60773066 & 60736014) and National Hi-tech Program (Project No.2006AA010108), and the Natural Scientific Reserach Innovation Foundation in Harbin Institute of Technology (Grant No. HIT.NSFIR.20009070).

References

Albrecht S. Joshua and Rebecca Hwa. 2007. *A Reexamination of Machine Learning Approaches for Sentence-Level MT Evaluation*. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 880-887.

Amigó Enrique, Julio Gonzalo, Anselmo Pènas, and Felisa Verdejo. 2005. *QARLA: a Framework for the Evaluation of Automatic Summarization.* In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics.

Amigó Enrique, Jesús Giménez, Julio Gonzalo, Felisa Verdejo. 2009. *The Contribution of Linguistic Features to Automatic Machine Translation Evaluation.* In 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.

Amigó Enrique, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. *MT Evaluation: Human-Like vs. Human Acceptable.* In Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistic, pages 17–24.

Banerjee Satanjeev and Alon Lavie. 2005. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures.

Blatz John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. *Confidence estimation for machine translation.* In Technical Report Natural Language Engineering Workshop Final Report, pages 97-100.

Callison-Burch Chris, Miles Osborne, and Philipp Koehn. 2006. *Re-evaluating the Role of BLEU in Machine Translation Research.* In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics

Chan S. Yee and Hwee T. Ng. 2008. *MAXSIM: A maximum similarity metric for machine translation evaluation.* In Proceedings of ACL-08: HLT, pages 55–62.

Doddington George. 2002. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.* In Proceedings of the 2nd International Conference on Human Language Technology, pages 138–145.

Drucker Harris, Chris J. C. Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik. 1996. *Support vector regression machines.* In NIPS.

Giménez Jesús and Lluís Màrquez. 2007. *Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems.* In Proceedings of the ACL Workshop on Statistical Machine Translation.

Giménez Jesús and Lluís Màrquez. 2008a. *Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations.* In Proceedings of IJCNLP, pages 319–326.

Giménez Jesús and Lluís Màrquez. 2008b. *On the Robustness of Linguistic Features for Automatic MT Evaluation.*

Joachims Thorsten. 2002. *Optimizing search engines using clickthrough data.* In KDD.

Klein Dan and Christopher D. Manning. 2003a. *Fast Exact Inference with a Factored Model for Natural Language Parsing.* In Advances in Neural Information Processing Systems 15, pp. 3-10.

Klein Dan and Christopher D. Manning. 2003b. *Accurate Unlexicalized Parsing.* Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Le Audrey and Mark Przybocki. 2005. *NIST 2005 machine translation evaluation official results.* In Official release of automatic evaluation scores for all submission.

Lin Chin-Yew and Franz Josef Och. 2004. *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics.* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 605-612.

Liu Ding and Daniel Gildea. 2005. *Syntactic Features for Evaluation of Machine Translation.* In Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 25–32.

Liu Ding and Daniel Gildea. 2007. *Source Language Features and Maximum Correlation Training for Machine Translation Evaluation.* In proceedings of NAACL HLT 2007, pages 41–48

Mehay Dennis and Chris Brew. 2007. *BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation.* In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation.

Melamed Dan I., Ryan Green, and Joseph P. Turian. 2003. *Precision and Recall of Machine Translation.* In Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics.

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

Owczarzak Karolina, Declan Groves, Josef Van Genabith, and Andy Way. 2006. *Contextual Bitext- Derived Paraphrases in Automatic MT Evaluation.* In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 148–155.

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

Papineni Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics.

Popović Maja and Hermann Ney. 2007. *Word Error Rates: Decomposition over POS classes and Applications for Error Analysis.* In Proceedings of the Second Workshop on Statistical Machine Translation, pages 48–55.

Popović Maja and Hermann Ney. 2009. *Syntax-oriented evaluation measures for machine translation output.* In Proceedings of the 4th EACL Workshop on Statistical Machine Translation, pages 29–32.

Snover Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. *A study of translation edit rate with targeted human annotation.* In Proceedings of AMTA, pages 223–231.

Tillmann Christoph, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. 1997. *Accelerated DP based Search for Statistical Translation.* In Proceedings of European Conference on Speech Communication and Technology.