MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric

Chi-kiu LO and **Dekai WU**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo|dekai}@cs.ust.hk

Abstract

The linguistically transparent MEANT and UMEANT metrics are tunable, simple yet highly effective, fully automatic approximation to the human HMEANT MT evaluation metric which measures semantic frame similarity between MT output and reference translations. In this paper, we describe HKUST's submission to the WMT 2013 metrics evaluation task, MEANT and UMEANT. MEANT is optimized by tuning a small number of weights—one for each semantic role label—so as to maximize correlation with human adequacy judgment on a development set. UMEANT is an unsupervised version where weights for each semantic role label are estimated via an inexpensive unsupervised approach, as opposed to MEANT's supervised method relying on more expensive grid search. In this paper, we present a battery of experiments for optimizing MEANT on different development sets to determine the set of weights that maximize MEANT's accuracy and stability. Evaluated on test sets from the WMT 2012/2011 metrics evaluation, both MEANT and UMEANT achieve competitive correlations with human judgments using nothing more than a monolingual corpus and an automatic shallow semantic parser.

1 Introduction

We evaluate in the context of WMT 2013 the MEANT (Lo *et al.*, 2012) and UMEANT (Lo and Wu, 2012) semantic machine translation (MT) evaluation metrics—tunable, simple yet highly effective, fully-automatic semantic frame based objective functions that score the degree of similarity between the MT output and the reference translations via semantic role labels (SRL). Recent studies (Lo *et al.*, 2013; Lo and Wu, 2013) show that tuning MT systems against MEANT more robustly improves translation adequacy, compared to tuning against BLEU or TER.

In the past decade, the progress of machine translation (MT) research is predominantly driven by the fast and cheap n-gram based MT evaluation metrics, such as BLEU (Papineni *et al.*, 2002), which assume that a good translation is one that shares the same lexical choices as the reference translation. Despite enforcing fluency, it has been established that these metrics do not enforce translation utility adequately and often fail to preserve meaning closely (Callison-Burch *et al.*, 2006; Koehn and Monz, 2006). Unlike BLEU, or other n-gram based MT evaluation metrics, MEANT adopts at outset the principle that a good translation is one from which the human readers may successfully understand at least the central meaning of the input sentence as captured by the basic event structure—"*who did what to whom, when, where and why"*(Pradhan *et al.*, 2004).

Lo *et al.* (2012) show that MEANT correlates better with human adequacy judgment than other commonly used automatic MT evaluation metrics, such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006). Recent studies (Lo *et al.*, 2013; Lo and Wu, 2013) also show that tuning MT system against MEANT produces more robustly adequate translations on both formal news text genre and informal web forum or public speech genre compared to tuning against BLEU or TER. These studies show that MEANT is a tunable and highly-accurate MT evaluation metric that drives MT system development towards higher utility.

As described in Lo and Wu (2011a), the pa-

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 422–428,
Sofia, Bulgaria, August 8-9, 2013

rameters in MEANT, i.e. the weight for each semantic role label, could be estimated using simple grid search to optimize the correlation with human adequacy judgments. Later, Lo and Wu (2012) described an unsupervised approach for estimating the parameters of MEANT using relative frequency of each semantic role label in the reference translations under the situation when the human judgments for the development set are unavailable. In this paper, we refer the version of MEANT using the unsupervised approach of weight estimation as UMEANT.

In this paper, we present a battery of experiments for optimizing MEANT on different development sets to determine the set of weights that maximizes MEANT's accuracy and stability. Evaluated on the test sets of WMT 2012/2011 metrics evaluation, MEANT and UMEANT achieve a competitive correlation score with human judgments by nothing more than a monolingual corpus and an automatic shallow semantic parser.

2 Related work

2.1 Lexical similarity based metrics

N-gram or edit distance based metrics such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) do not correctly reflect the similarity of the basic event structure— "*who did what to whom, when, where and why*"— of the input sentence. In fact, a number of large scale meta-evaluations (Callison-Burch *et al.*, 2006; Koehn and Monz, 2006) report cases where BLEU strongly disagrees with human judgments of translation adequacy.

Although AMBER (Chen *et al.*, 2012) shows a high correlation with human adequacy judgment (Callison-Burch et al., 2012) and claims to preserve the simplicity of BLEU, the modifications it incurred on BLEU through four different n-gram matching strategies and several different penalties makes it very hard to interpret and indicate what errors the MT systems are making.

2.2 Linguistic feature based metrics

ULC (Giménez and Màrquez, 2007, 2008) is an automatic metric that incorporates several semantic similarity features and shows improved correlation with human judgement of translation quality (Callison-Burch *et al.*, 2007; Giménez and Màrquez, 2007; Callison-Burch *et al.*, 2008; Giménez and Màrquez, 2008) but no work has been done towards tuning an SMT system using a pure form of ULC perhaps due to its expensive run time. Lambert *et al.* (2006) did tune on QUEEN, a simplified version of ULC that discards the semantic features of ULC and is based on pure lexical similarity. Therefore, QUEEN suffers from the problem of failing to reflect translation adequacy similar to other n-gram based metrics.

Similarly, SPEDE (Wang and Manning, 2012) is an integrated probabilistic FSM and probabilistic PDA model that predicts the edit sequence needed for the MT output to match the reference. Sagan (Castillo and Estrella, 2012) is a semantic textual similarity metric based on a complex textual entailment pipeline. These aggregated metrics require sophisticated feature extraction steps; contain several dozens of parameters to tune and employ expensive linguistic resources, like WordNet and paraphrase table. Like ULC, these matrices are not useful in the MT system development cycle for tuning due to expensive running time. The metrics themselves are also expensive in training and tuning due to the large number of parameters to be estimated. Although ROSE (Song and Cohn, 2011) is a weighted linear model of shallow linguistic features which is cheaper in run time but it still contains several dozens of weights that need to be tuned which affects the portability of the metric for evaluating translations across domains.

Rios *et al.* (2011) introduced TINE, an automatic recall-oriented evaluation metric which aims to preserve the basic event structure, but no work has been done toward tuning an SMT system against it. TINE performs comparably to BLEU and worse than METEOR on correlation with human adequacy judgment.

3 MEANT and UMEANT

MEANT (Lo *et al.*, 2012), which is the weighted f-measure over the matched semantic role labels of the automatically aligned semantic frames and role fillers, outperforms BLEU, NIST, METEOR, WER, CDER and TER. Recent studies (Lo *et al.*, 2013; Lo and Wu, 2013) also show that tuning MT system against MEANT produces more robustly adequate translations than the common practice of tuning against BLEU or TER across different data genres, such as formal newswire text, informal web forum text and informal public speech. Pre-

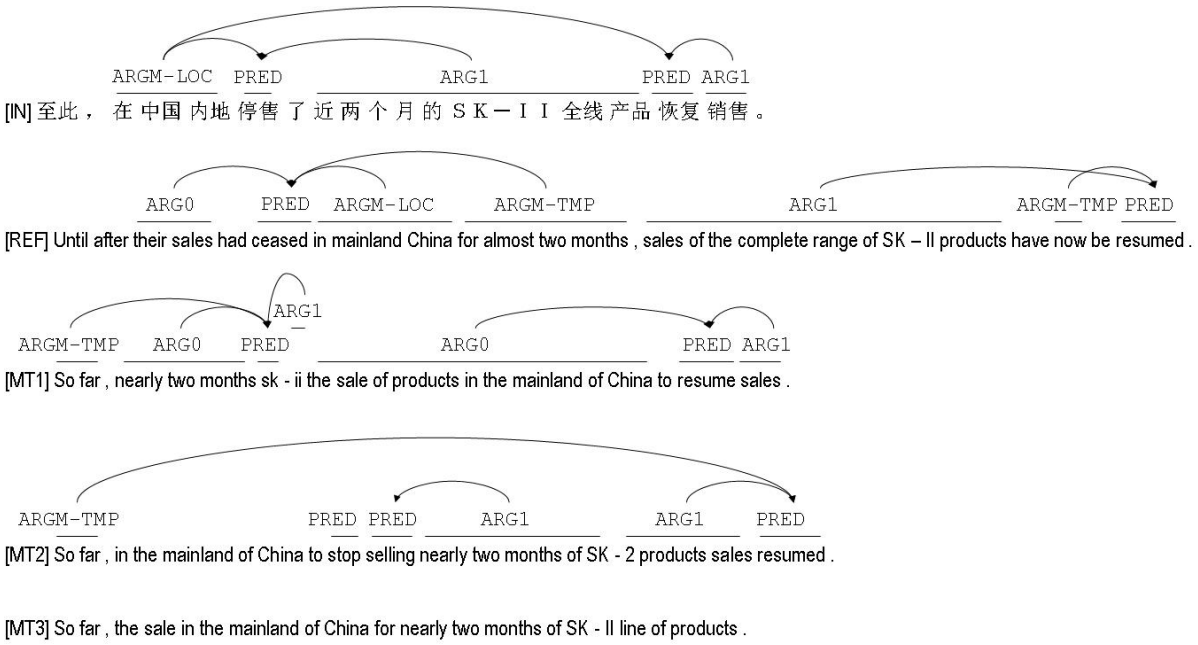

Figure 1: Examples of automatic shallow semantic parses. The input is parsed by a Chinese automatic shallow semantic parser. The reference and MT output are parsed by an English automatic shallow semantic parser. There are no semantic frames for MT3 since there is no predicate.

cisely, MEANT is computed as follows:

1. Apply an automatic shallow semantic parser on both the references and MT output. (Figure 1 shows examples of automatic shallow semantic parses on both reference and MT output.)

2. Apply maximum weighted bipartite matching algorithm to align the semantic frames between the references and MT output by the lexical similarity of the predicates.

3. For each pair of aligned semantic frames,

   (a) Lexical similarity scores determine the similarity of the semantic role fillers.

   (b) Apply maximum weighted bipartite matching algorithm to align the semantic role fillers between the reference and MT output according to their lexical similarity.

4. Compute the weighted f-measure over the matching role labels of these aligned predicates and role fillers.

$$
\begin{aligned}
M_{i,j} &\equiv \text{total \# ARG j of aligned frame i in MT} \\
R_{i,j} &\equiv \text{total \# ARG j of aligned frame i in REF} \\
S_{i,\text{pred}} &\equiv \text{similarity of predicate in aligned frame i} \\
S_{i,j} &\equiv \text{similarity of ARG j in aligned frame i} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j}
\end{aligned}
$$

$$
m_i \equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}
$$

$$
r_i \equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}
$$

$$
\text{precision} = \frac{\sum_i m_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j M_{i,j}}}{\sum_i m_i}
$$

$$
\text{recall} = \frac{\sum_i r_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j R_{i,j}}}{\sum_i r_i}
$$

where m_{i} and r_{i} are the weights for frame, i, in the MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. $M_{\text{i,j}}$ and $R_{\text{i,j}}$ are the total counts of argument of type j in frame i in the MT and REF respectively. $S_{i,\text{pred}}$ and $S_{i,j}$ are the lexical similarities of the predicates and role fillers of the arguments of type j between the reference translations and the MT output. w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type$j$ between the reference translations and the MT output. There are in total 12 weights for the set of

semantic role labels in MEANT as defined in Lo and Wu (2011b).

For MEANT, w_{pred} and w_j are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a). For UMEANT, w_{pred} and w_j are estimated in an unsupervised manner using relative frequency of each semantic role label in the reference translations when the human judgments on adequacy of the development set were unavailable (Lo and Wu, 2012).

In this experiment, we use a MEANT / UMEANT implementation along the lines described in Lo *et al.* (2012) and Tumuluru *et al.* (2012) but we incorporate a variant of the aggregation function proposed in Mihalcea *et al.* (2006) for phrasal similarity of role fillers as it normalizes the phrase length better than geometric mean as described in Tumuluru *et al.* (2012). In case there is no semantic frame in the sentence, we treat the whole sentence as a phrase and calculate the phrasal similarity, like the role fillers in step 3.1, as the MEANT score.

4 Experimental setup

We tune the 12 weights for the set of semantic role labels in MEANT using grid search to maximize the correlation with human judgment on 6 development sets. Following the protocol in WMT12 metrics evaluation task (Callison-Burch *et al.*, 2012), we use Kendall's correlation coefficient for the sentence-level correlation with human judgments.

The GALE development set consists of 40 sentences randomly drawn from the DARPA GALE P2.5 Chinese-English evaluation set along with the outputs from 3 participating MT systems and the corresponding human adequacy judgments. The WMT12-A development set consists of 800 sentences randomly drawn from the Czech-English test set in WMT12 metrics evaluation task along with the output from 5 participating systems and the corresponding human judgments. Similarly, each of the WMT12-B, WMT12-C and WMT12-D development sets consists of 800 randomly drawn sentences from the WMT12 metrics evaluation test set on German-English, Spanish-English and French-English respectively. The WMT12-E development set consists of 800 sentences out of which 200 sentences were randomly drawn from each of WMT12-A, WMT12-B, WMT12-C and WMT12-D data set.

We evaluated MEANT and UMEANT on 3 groups of test sets. The first group is the original (without partition) test data for each language pair (translated in English) in WMT12. This group of test sets is used for comparing MEANT's performance with the reported results from other participants of WMT12. The second group is the held out subset of the test data for each language pair in WMT12. The third group is the original set of test data for each language pair in WMT11. The latter 2 groups are used for determining which set of tuned weights maximize the accuracy and stability of MEANT.

5 Results

Table 1 shows that the best and the worst sentence-level correlations reported in Callison-Burch *et al.* (2012) on the original WMT12 test sets (without partitioning) for translations into English, together the sentence-level correlation of MEANT tuned on different development sets and UMEANT. The grey boxes mark the results of experiments in which there was an overlap between parts of the development data and the test data. A study of the values for the 12 weights associated with the semantic role labels show that a general trend of the importance of different labels in MEANT: "who" is always the most important; "did", "what", "where", "why", "extent", "modal" and "other" are quite important too; "when", "manner" and "negation" fluctuate where they are quite important in some development sets but not quite important in some development sets; "whom" is usually not important. Given the fact that MEANT employs significantly less expensive linguistic resources and less sophisticated machine learning algorithm in tuning the parameters, the performance of MEANT is very competitive with other participants last year.

Table 2 shows the sentence-level correlation on the WMT12 held-out test sets and the original WMT11 test sets of MEANT tuned on different development sets and UMEANT together with the average sentence-level correlation on all test sets. The results show that MEANT tuning on WMT12-C development set achieve the highest sentence-level correlation with human judgments on average. UMEANT, the unsupervised wight estimated version of MEANT, achieves a very competitive correlation score when compared with MEANT tuned on different development sets. As a result,

Table 1: The best and the worst sentence-level correlation reported in Callison-Burch *et al.* (2012) on the original WMT12 test sets (without partitioning) for translations into English together the sentence-level correlation of MEANT tuned on different development sets and UMEANT. The grey box marked results of experiments in which parts of the development data and the test data are overlapped.

	WMT12 cz-en	WMT12 de-en	WMT12 es-en	WMT12 fr-en
Best reported	0.21	0.28	0.26	0.26
MEANT (GALE)	0.13	0.16	0.15	0.15
MEANT (WMT12-A)	0.12	0.17	0.16	0.15
MEANT (WMT12-B)	0.11	0.18	0.15	0.14
MEANT (WMT12-C)	0.12	0.17	0.17	0.15
MEANT (WMT12-D)	0.12	0.17	0.16	0.16
MEANT (WMT12-E)	0.12	0.17	0.17	0.15
UMEANT	0.12	0.17	0.16	0.14
Worst reported	0.06	0.08	0.08	0.07

Table 2: Sentence-level correlation on the WMT12 held-out test sets and the original WMT11 test sets of MEANT tuned on different development sets and UMEANT together with the average sentence-level correlation on all test sets.

	WMT12 held-out				WMT11				Average
	cz-en	de-en	es-en	fr-en	cz-en	de-en	es-en	fr-en	-
MEANT (GALE)	0.0657	0.1251	0.1762	**0.1719**	0.3460	0.1123	0.2416	0.1913	0.1788
MEANT (WMT12-A)	0.0652	0.1117	0.1663	0.1540	0.3764	0.1101	0.2314	0.1944	0.1762
MEANT (WMT12-B)	0.0458	0.1294	0.1556	0.1548	**0.3992**	**0.1479**	0.2571	**0.2037**	0.1867
MEANT (WMT12-C)	**0.0746**	0.1278	**0.1833**	0.1592	0.3764	0.1324	0.2674	0.1882	**0.1887**
MEANT (WMT12-D)	0.0628	0.1164	0.1826	0.1655	0.3802	0.1168	0.2339	0.1975	0.1820
MEANT (WMT12-E)	0.0496	**0.1353**	0.1791	0.1619	0.3840	0.1101	0.2596	0.1851	0.1831
UMEANT	0.0477	0.1333	0.1606	0.1548	0.3764	0.1257	**0.2828**	0.1913	0.1841

we submitted two metrics to WMT 2013 metrics evaluation task. One is MEANT with weights learned from tuning on WMT12-C development sets and the other submission is UMEANT.

6 Conclusion

In this paper, we have evaluated in the context of WMT2013 the MEANT and UMEANT metrics, which are tunable, accurate yet inexpensive fully automatic machine translation evaluation metrics that measure similarity between the MT output and the reference via semantic frames. Recent studies show that tuning MT system against MEANT produces more robustly adequate translations than the common practice of tuning against BLEU or TER across different data genres, such as formal newswire text, informal web forum text and informal public speech. The weight for each semantic role label in MEANT is estimated by maximizing the correlation with human adequacy judgment on a development set. UMEANT is a version of MEANT in which weight for each semantic role label is estimated in an unsupervised fashion using the relative frequency of the semantic role labels in the reference. We present the experimental results for determining the set of weights that maximize MEANT's accuracy and stability by optimizing MEANT on different development sets.

We disagree with the notion "a good evaluation metric is not necessarily a good tuning metric, and vice versa" (Chen *et al.*, 2012). Instead, we believe that a good evaluation metric should be one that is a good objective function to drive the development of MT systems towards higher utility. In other words, a good evaluation metric should correlate well with human adequacy judgment and at the same time, be inexpensive in running time so as to fit into the MT pipeline to improve MT quality. Our results shows that MEANT is a good evaluation/tuning metric because it achieves a competitive correlation score with human judgments by using less expensive linguistic resources and training algorithms making it possible to tune MT system against MEANT to improve MT quality.

7 Acknowledgment

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agree-

ment no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

References

Satanjeev Banerjee and Alon Lavie. 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 for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in Machine Translation Research. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 249–256, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 136–158, 2007.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-evaluation of Machine Translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 70–106, 2008.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 10–51, 2012.

Julio Castillo and Paula Estrella. Semantic Textual Similarity for MT evaluation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 52–58, 2012.

Boxing Chen, Roland Kuhn, and George Foster. Improving AMBER, an MT Evaluation Metric. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 59–63, 2012.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 138–145, San Diego, California, 2002.

Jesús Giménez and Lluís Màrquez. 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 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June 2008.

Philipp Koehn and Christof Monz. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation (WMT-06)*, pages 102–121, 2006.

Patrik Lambert, Jesús Giménez, Marta R Costa-jussá, Enrique Amigó, Rafael E Banchs, Lluıs Márquez, and JAR Fonollosa. Machine Translation system development based on human likeness. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 246–249. IEEE, 2006.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 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)*, 2006.

Chi-kiu Lo and Dekai Wu. MEANT: An Inexpensive, High-Accuracy, Semi-Automatic Metric for Evaluating Translation Utility based on Semantic Roles. In *Proceedings of the Joint conference of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT-11)*, 2011.

Chi-kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Proceedings of the 6th Workshop on Syntax and Structure in Statistical Translation (SSST-6)*, 2012.

Chi-kiu Lo and Dekai Wu. Can informal genres be better translated by tuning on automatic semantic metrics? In *Proceedings of the 14th Machine Translation Summit (MTSummit-XIV)*, 2013.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully Automatic Semantic MT Evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT2012)*, 2012.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*, 2013.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 775. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 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)*, 2000.

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

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, 2004.

Miguel Rios, Wilker Aziz, and Lucia Specia. Tine: A metric to assess MT adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-2011)*, pages 116–122, 2011.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Xingyi Song and Trevor Cohn. Regression and Ranking based Optimisation for Sentence Level Machine Translation Evaluation. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT 2011)*, pages 123–129, 2011.

Anand Karthik Tumuluru, Chi-kiu Lo, and Dekai Wu. Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic mt evaluation. In *Proceeding of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC-26)*, 2012.

Mengqiu Wang and Christopher D. Manning. SPEDE: Probabilistic Edit Distance Metrics for MT Evaluation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 76–83, 2012.