# Improving machine translation by training against an automatic semantic frame based evaluation metric

**Chi-kiu Lo** and **Karteek Addanki** and **Markus Saers** and **Dekai Wu**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo|vskaddanki|masaers|dekai}@cs.ust.hk

## Abstract

We present the first ever results showing that tuning a machine translation system against a semantic frame based objective function, MEANT, produces more robustly adequate translations than tuning against BLEU or TER as measured across commonly used metrics and human subjective evaluation. Moreover, for informal web forum data, human evaluators preferred MEANT-tuned systems over BLEU- or TER-tuned systems by a significantly wider margin than that for formal newswire—even though automatic semantic parsing might be expected to fare worse on informal language. We argue that by preserving the meaning of the translations as captured by semantic frames right in the training process, an MT system is constrained to make more accurate choices of both lexical and reordering rules. As a result, MT systems tuned against semantic frame based MT evaluation metrics produce output that is more adequate. Tuning a machine translation system against a semantic frame based objective function is independent of the translation model paradigm, so, any translation model can benefit from the semantic knowledge incorporated to improve translation adequacy through our approach.

## 1 Introduction

We present the first ever results of tuning a statistical machine translation (SMT) system against a semantic frame based objective function in order to produce a more adequate output. We compare the performance of our system with that of two baseline SMT systems tuned against BLEU and TER, the commonly used n-gram and edit distance based metrics. Our system performs better than the baseline across seven commonly used evaluation metrics and subjective human evaluation on adequacy. Surprisingly, tuning against a semantic MT evaluation metric also significantly outperforms the baseline on the domain of informal web forum data wherein automatic semantic parsing might be expected to fare worse. These results strongly indicate that using a semantic frame based objective function for tuning would drive development of MT towards direction of higher utility.

Glaring errors caused by semantic role confusion that plague the state-of-the-art MT systems are a consequence of using fast and cheap lexical n-gram based objective functions like BLEU to drive their development. 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).

We argue that instead of BLEU, a metric that focuses on getting the meaning right should be used as an objective function for tuning SMT so as to drive continuing progress towards higher utility. MEANT (Lo *et al.*, 2012), is an automatic semantic MT evaluation metric that measures similarity between the MT output and the reference translation via semantic frames. It correlates better with human adequacy judgment than other automatic MT evaluation metrics. Since a high MEANT score is contingent on correct lexical choices as well as syntactic and semantic structures, we believe that tuning against MEANT would improve both translation adequacy and fluency.

Incorporating semantic structures into SMT by tuning against a semantic frame based evaluation metric is independent of the MT paradigm. Therefore, systems from different MT paradigms (such as hierarchical, phrase based, transduction grammar based) can benefit from the semantic information incorporated through our approach.



*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 375–381,
Sofia, Bulgaria, August 4-9 2013. 

## 2 Related Work

Relatively little work has been done towards biasing the translation decisions of an SMT system to produce adequate translations that correctly preserve *who did what to whom, when, where and why* (Pradhan *et al.*, 2004). This is because the development of SMT systems was predominantly driven by tuning against n-gram based evaluation metrics such as BLEU or edit distance based metrics such as TER which do not sufficiently bias SMT system's decisions to produce adequate translations. Although there has been a recent surge of work aimed towards incorporating semantics into the SMT pipeline, none attempt to tune against a semantic objective function. Below, we describe some of the attempts to incorporate semantic information into the SMT and present a brief survey on evaluation metrics that focus on rewarding semantically valid translations.

**Utilizing semantics in SMT** In the past few years, there has been a surge of work aimed at incorporating semantics into various stages of the SMT. Wu and Fung (2009) propose a two-pass model that reorders the MT output to match the SRL of the input, which is too late to affect the translation decisions made by the MT system during decoding. In contrast, training against a semantic objective function attempts to improve the decoding search strategy by incorporating a bias towards meaningful translations into the model instead of postprocessing its results.

Komachi *et al.* (2006) and Wu *et al.* (2011) preprocess the input sentence to match the verb frame alternations in the output side. Liu and Gildea (2010) and Aziz *et al.* (2011) use input side SRL to train a tree-to-string SMT system. Xiong *et al.* (2012) trained a discriminative model to predict the position of the semantic roles in the output. All these approaches are orthogonal to the present question of whether to train toward a semantic objective function. Any of the above models could potentially benefit from tuning with semantic metrics.

**MT evaluation metrics** As mentioned previously, tuning against n-gram based metrics such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) does not sufficiently drive SMT into making decisions to produce adequate translations that correctly preserve *"who did what to whom, when, where and why"*. 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 accuracy. Tuning against edit distance based metrics such as CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) also fails to sufficiently bias SMT systems towards producing translations that preserve semantic information.

We argue that an SMT system tuned against an adequacy-oriented metric that correlates well with human adequacy judgement produces more adequate translations. For this purpose, we choose MEANT, an automatic semantic MT evaluation metric that focuses on getting the meaning right by comparing the semantic structures of the MT output and the reference. We briefly describe some of the alternative semantic metrics below to justify our choice.

ULC (Giménez and Màrquez, 2007, 2008) is an aggregated metric that incorporates several semantic similarity features and shows improved correlation with human judgement on 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 MT system against ULC perhaps due to its expensive running time. Lambert *et al.* (2006) did tune on QUEEN, a simplified version of ULC that discards the semantic features and is based on pure lexical features. Although tuning on QUEEN produced slightly more preferable translations than solely tuning on BLEU, the metric does not make use of any semantic features and thus fails to exploit any potential gains from tuning to semantic objectives.

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

In contrast to TINE, MEANT (Lo *et al.*, 2012), which is the weighted f-score over the matched semantic role labels of the automatically aligned semantic frames and role fillers, outperforms BLEU, NIST, METEOR, WER, CDER and TER. This makes it more suitable for tuning SMT systems to produce much adequate translations.

| newswire | BLEU | NIST | METEOR no_syn | METEOR | WER | CDER | TER | MEANT |
|---|---|---|---|---|---|---|---|---|
| BLEU-tuned | 29.85 | 8.84 | 52.10 | 55.42 | 67.88 | 55.67 | **58.40** | 0.1667 |
| TER-tuned | 25.37 | 6.56 | 48.26 | 51.24 | 66.18 | 52.58 | 56.96 | 0.1578 |
| MEANT-tuned | **25.91** | **7.81** | **50.15** | **53.60** | **67.76** | **54.56** | 58.61 | **0.1676** |

Table 1: Translation quality of MT system tuned against MEANT, BLEU and TER on newswire data

| forum | BLEU | NIST | METEOR no_syn | METEOR | WER | CDER | TER | MEANT |
|---|---|---|---|---|---|---|---|---|
| BLEU-tuned | 9.58 | 4.10 | 31.77 | 34.63 | 80.09 | 64.54 | 76.12 | 0.1711 |
| TER-tuned | 6.94 | 2.21 | 28.55 | 30.85 | 76.15 | 57.96 | 74.73 | 0.1539 |
| MEANT-tuned | **7.92** | **3.11** | **30.40** | **33.08** | **77.32** | **61.01** | **74.64** | **0.1727** |

Table 2: Translation quality of MT system tuned against MEANT, BLEU and TER on forum data

## 3 Tuning SMT against MEANT

We now show that using MEANT as an objective function to drive minimum error rate training (MERT) of state-of-the-art MT systems improves MT utility not only on formal newswire text, but even on informal forum text, where automatic semantic parsing is difficult.

Toward improving translation utility of state-of-the-art MT systems, we chose to use a strong and competitive system in the DARPA BOLT program as our baseline. The baseline system is a Moses hierarchical model trained on a collection of LDC newswire and a small portion of Chinese-English parallel web forum data, together with a 5-gram language model. For the newswire experiment, we used a collection of NIST 02-06 test sets as our development set and NIST 08 test set for evaluation. The development and test sets contain 6,331 and 1,357 sentences respectively with four references. For the forum data experiment, the development and test sets were a held-out subset of the BOLT phase 1 training data. The development and test sets contain 2,000 sentences and 1,697 sentences with one reference.

We use ZMERT (Zaidan, 2009) to tune the baseline because it is a widely used, highly competitive, robust, and reliable implementation of MERT that is also fully configurable and extensible with regard to incorporating new evaluation metrics. In this experiment, we use a MEANT implementation along the lines described in Lo *et al.* (2012).

In each experiment, we tune two contrastive conventional 100-best MERT tuned baseline systems on both newswire and forum data genres; one tuned against BLEU, an n-gram based evaluation metric and the other using TER, an edit distance based metric. As semantic role labeling is expensive we only tuned using 10-best list for MEANT-tuned system. Tuning against BLEU and TER took around 1.5 hours and 5 hours per iteration respectively whereas tuning against MEANT took about 1.6 hours per iteration.

## 4 Results

Of course, tuning against any metric would maximize the performance of the SMT system on that particular metric, but would be overfitting. For example, something would be seriously wrong if tuning against BLEU did not yield the best BLEU scores. A far more worthwhile goal would be to bias the SMT system to produce adequate translations while achieving the best scores across all the metrics. With this as our objective, we present the results of comparing MEANT-tuned systems against the baselines as evaluated on commonly used automatic metrics and human adequacy judgement.

**Cross-evaluation using automatic metrics** Tables 1 and 2 show that MEANT-tuned systems achieve the best scores across all other metrics in both newswire and forum data genres, when avoiding comparison of the overfit metrics too similar to the one the system was tuned on (the cells shaded in grey in the table: NIST and METEOR are n-gram based metrics, similar to BLEU while WER and CDER are edit distance based metrics, similar to TER). In the newswire domain, however, our system achieves marginally lower TER score than BLEU-tuned system.

Figure 1 shows an example where the MEANT-tuned system produced a more adequate translation that accurately preserves the semantic structure of the input sentence than the two baseline systems. The MEANT scores for the MT output from the BLEU-, TER- and MEANT-tuned systems are 0.0635, 0.1131 and 0.2426 respectively. Both the MEANT score and the human evaluators rank the MT output from the MEANT-tuned sys-

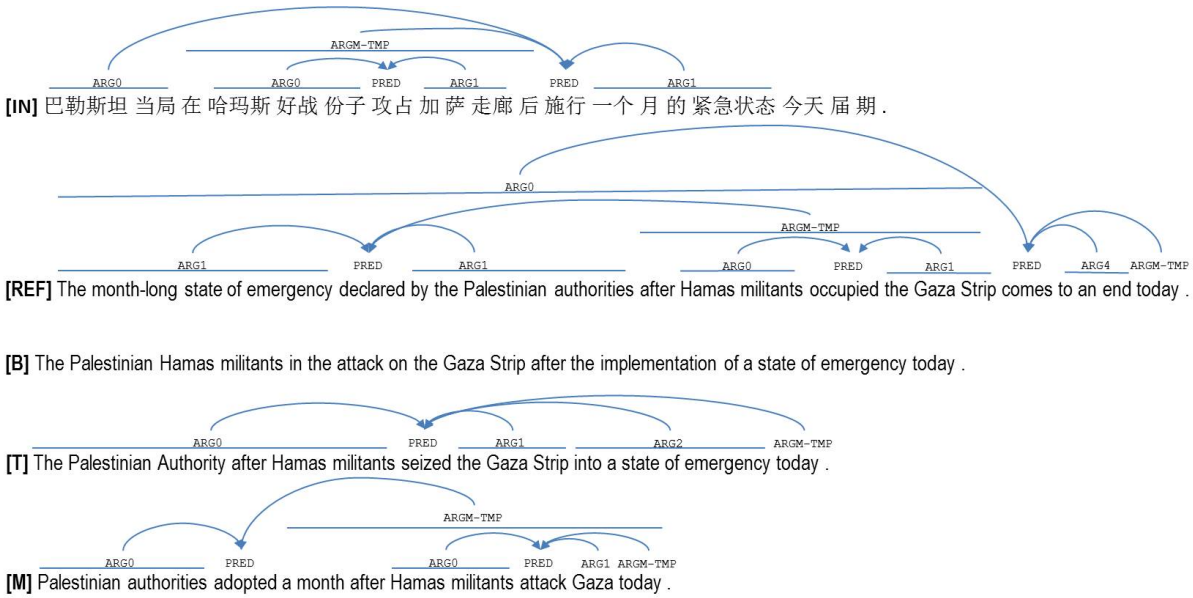


Figure 1: Examples of machine translation output and the corresponding semantic parses from the **[B]** BLEU-, **[T]** TER-and **[M]** MEANT-tuned systems together with **[IN]** the input sentence and **[REF]** the reference translation. Note that the MT output of the BLEU-tuned system has no semantic parse output by the automatic shallow semantic parser.

tem as the most adequate translation. In this example, the MEANT-tuned system has translated the two predicates "攻占" and "施行" in the input sentence into the correct form of the predicates "attack" and "adopted" in the MT output, whereas the BLEU-tuned system has translated both of them incorrectly (translates the predicates into nouns) and the TER-tuned system has correctly translated only the first predicate (into "seized") and dropped the second predicate. Moreover, for the frame "攻占" in the input sentence, the MEANT-tuned system has correctly translated the ARG0 "哈玛斯好战份子" into "Hamas militants" and the ARG1 "加萨走廊" into "Gaza". However, the TER-tuned system has dropped the predicate "施行" so that the corresponding arguments "The Palestinian Authority" and "into a state of emergency" have all been incorrectly associated with the predicate "攻占/seized". This example shows that the translation adequacy of SMT has been improved by tuning against MEANT because the MEANT-tuned system is more accurately preserving the semantic structure of the input sentence.

Our results show that MEANT-tuned system maintains a balance between lexical choices and word order because it performs well on n-gram based metrics that reward lexical matching and edit distance metrics that penalize incorrect word order. This is not surprising as a high MEANT score relies on a high degree of semantic structure matching, which is contingent upon correct lexical choices as well as syntactic and semantic structures.

**Human subjective evaluation** In line with our original objective of biasing SMT systems towards producing adequate translations, we conduct a human evaluation to judge the translation utility of the outputs produced by MEANT-, BLEU- and TER-tuned systems. Following the manual evaluation protocol of Lambert *et al.* (2006), we randomly draw 150 sentences from the test set in each domain to form the manual evaluation set. Table 3 shows the MEANT scores of the two manual evaluation sets. In both evaluation sets, like in the test sets, the output from the MEANT-tuned system score slightly higher in MEANT than that from the BLEU-tuned system and significantly higher than that from the TER-tuned system. The output of each tuned MT system along the input sentence and the reference were presented to human evaluators. Each evaluation set is ranked by two evaluators for measuring inter-evaluator agreement.

Table 4 indicates that output of the MEANT-tuned system is ranked adequate more frequently compared to BLEU- and TER-tuned baselines for both newswire and web forum genres. The inter-

| | newswire | forum |
|---|---|---|
| BLEU-tuned | 0.1564 | 0.1663 |
| TER-tuned | 0.1203 | 0.1453 |
| MEANT-tuned | 0.1633 | 0.1737 |

Table 3: MEANT scores of each system in the 150-sentence manual evaluation set.

| | newswire | | forum | |
|---|---|---|---|---|
| | Eval 1 | Eval 2 | Eval 1 | Eval 2 |
| BLEU-tuned (B) | 37 | 42 | 47 | 42 |
| TER-tuned (T) | 22 | 24 | 28 | 23 |
| MEANT-tuned (M) | 55 | 56 | 59 | 68 |
| B=T | 14 | 12 | 0 | 0 |
| M=B | 5 | 4 | 8 | 9 |
| M=T | 4 | 4 | 4 | 4 |
| M=B=T | 13 | 9 | 4 | 4 |

Table 4: No. of sentences ranked the most adequate by human evaluators for each system.

| $H_1$ | newswire | forum |
|---|---|---|
| MEANT-tuned > BLEU-tuned | 80% | 95% |
| MEANT-tuned > TER-tuned | 99% | 99% |

Table 5: Significance level of accepting the alternative hypothesis.

evaluator agreement is 84% and 70% for newswire and forum data genres respectively.

We performed the right-tailed two proportion significance test on human evaluation of the SMT system outputs for both the genres. Table 5 shows that the MEANT-tuned system generates more adequate translations than the TER-tuned system at the 99% significance level for both newswire and web forum genres. The MEANT-tuned system is ranked more adequate than the BLEU-tuned system at the 95% significance level on the web forum genre and for the newswire genre the hypothesis is accepted at a significance level of 80%. The high inter-evaluator agreement and the significance tests confirm that MEANT-tuned system is better at producing adequate translations compared to BLEU- or TER-tuned systems.

**Informal vs. formal text** The results of table 4 and 5 also show that—surprisingly—the human evaluators preferred MEANT-tuned system output over BLEU-tuned and TER-tuned system output by a far wider margin on the informal forum text compared to the formal newswire text. The MEANT-tuned system is better than both baselines at the 80% significance level for the formal text genre. For the informal text genre, it performs the two baselines at the 95% significance level. Although one might expect an semantic frame dependent metric such as MEANT to perform poorly on the domain of informal text, surprisingly, it nonetheless significantly outperforms the baselines at the task of generating adequate output. This indicates that the design of the MEANT evaluation metric is robust enough to tune an SMT system towards adequate output on informal text domains despite the shortcomings of automatic shallow semantic parsing.

## 5 Conclusion

We presented the first ever results to demonstrate that tuning an SMT system against MEANT produces much adequate translation than tuning against BLEU or TER, as measured across all other commonly used metrics and human subjective evaluation. We also observed that tuning against MEANT succeeds in producing adequate output significantly more frequently even on the informal text such as web forum data. By preserving the meaning of the translations as captured by semantic frames right in the training process, an MT system is constrained to make more accurate choices of both lexical and reordering rules. The performance of our system as measured across all commonly used metrics indicate that tuning against a semantic MT evaluation metric does produce output which is adequate and fluent.

We believe that tuning on MEANT would prove equally useful for MT systems based on any paradigm, especially where the model does not incorporate semantic information to improve the adequacy of the translations produced and using MEANT as an objective function to tune SMT would drive sustainable development of MT towards the direction of higher utility.

## 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 agreement 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


Wilker Aziz, Miguel Rios, and Lucia Specia. Shallow semantic trees for SMT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT2011)*, 2011.

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.

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.

Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT 2006)*, 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.

Ding Liu and Daniel Gildea. Semantic role features for machine translation. In *Proceedings of the 23rd international conference on Computational Linguistics (COLING-10)*, 2010.

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.

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 *Proceed-*

*ings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122. Association for Computational Linguistics, 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.

Dekai Wu and Pascale Fung. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT-09)*, pages 13–16, 2009.

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Extracting preordering rules from predicate-argument structures. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*, 2011.

Deyi Xiong, Min Zhang, and Haizhou Li. Modeling the Translation of Predicate-Argument Structure for SMT. In *Proceedings of the Joint conference of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12)*, 2012.

Omar F. Zaidan. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88, 2009.