Can informal genres be better translated by tuning on automatic semantic metrics?

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

Even though the informal language of spoken text and web forum genres presents great difficulties for automatic semantic role labeling, we show that surprisingly, tuning statistical machine translation against the SRL-based objective function, MEANT, nevertheless leads more robustly to adequate translations of these informal genres than tuning against BLEU or TER. The accuracy of automatic semantic parsing has been shown to degrade significantly on informal genres such as speech or tweets, compared to formal genres like newswire. In spite of this, human evaluators preferred translations from MEANT-tuned systems over the BLEU- or TER-tuned ones by a significant margin. Error analysis indicates that one of the major sources of errors in automatic shallow semantic parsing of informal genres is failure to identify the semantic frame for copula or existential senses of "be". We show that MEANT's correlation with human adequacy judgment on informal text is improved by reconstructing the missing semantic frames for "be". Our tuning approach is independent of the translation model architecture, so any SMT model can potentially benefit from the semantic knowledge incorporated through our approach.

1 Introduction

We present the first known results of robustly improving the adequacy of statistical machine translation (SMT) for *informal* genres by tuning against the semantic frame based objective function MEANT. Using both spoken language and web forum datasets, we show that tuning against MEANT outperforms two SMT baselines that follow the common practice of tuning against the n-gram based BLEU metric or the edit distance based TER metric. We investigate results across a battery of automatic evaluation metrics, as well as subjective human evaluation of translation adequacy. Since automatic semantic parsing has been shown to fare worse on informal genres, where the robustness of the POS tagging and syntactic parsing that it depends on suffers, it is surprising that tuning against a semantic frame based objective function nonetheless performs more robustly than tuning against non-semantic metrics. Our results encouragingly suggest that further improvement of semantic frame based objective functions for training SMT will be a fruitful direction for raising the utility of machine translation on informal language, and not only formal text genres.

Previous work on improving machine translation of informal text has mostly focused on using domain adaptation techniques instead of incorporating semantics, because the accuracy of automatic shallow semantic parsers has been reported to drop by around 10% on speech data (Favre *et al.*, 2010) and by more than 30% on web data like tweets (Liu *et al.*, 2010). Yet something more must be done; common errors in machine translation of formal text caused by semantic role confusions of the kind that plague state-of-the-art MT systems are even more glaring for informal texts. Semantic role confusion errors in SMT are mainly the consequence of using fast and cheap lexical n-gram based objective functions such as BLEU to drive 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).

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 93–100.

We recently showed that the translation adequacy for formal news is improved by replacing the surface oriented metrics like BLEU or TER with a semantic objective function, when tuning the parameters of MT systems (Lo *et al.*, 2013). We now ask whether the same approach of tuning MT systems against a semantic objective function might improve translation adequacy even for informal texts, despite the fact that automatic semantic parsing is known to be of lower accuracy. If the objective function for tuning SMT sufficiently reflects preservation of meaning, this should ultimately drive continuing progress toward higher utility. Our fully-automatic, semantic MT evaluation metric, MEANT (Lo *et al.*, 2012), measures similarity between MT output and reference translations as judged via semantic frames. For formal text genres, we have shown that MEANT correlates more highly with human adequacy judgment than all other automatic MT evaluation metrics. Although MEANT has not been shown to work well for informal language genres where it is likely to suffer from degraded semantic parsing accuracy, since a high MEANT score is also contingent on correct lexical choices as well as syntactic and semantic structures, we hypothesize that perhaps tuning against MEANT would nonetheless improve both translation adequacy and fluency even on speech and web forum data.

The proposed approach of incorporating semantic structures into SMT by tuning the model against a semantic frame based evaluation metric is independent of assumptions about the underlying translation model architecture. Therefore, MT systems from different SMT approaches (such as hierarchical, phrase based, or synchronous/transduction grammar based) or those applying other techniques for informal data (such as domain adaptation from formal to informal text, or integration of linguistic features) could also benefit from the semantic information incorporated through our approach.

2 Related work

Relatively little work has been done toward addressing the problem of biasing the translation decisions of an SMT system to produce adequate translations for informal text that correctly preserve *who did what to whom, when, where and why* (Pradhan *et al.*, 2004). There has been a recent surge of work aimed at incorporating semantics into the SMT pipeline; however, none attempts to improve translation quality on *informal* text due to the difficulty of semantic parsing. Below, we describe some of the attempts to (a) improve informal text translation quality using domain adaptation techniques and (b) incorporate semantic role labeling information into the SMT pipeline and present a brief survey of evaluation metrics that focus on rewarding semantically valid translations.

2.1 Semantics in SMT

There is a rising trend of work aimed at incorporating semantics into various stages of the SMT pipeline, for example, preprocessing the input (Komachi *et al.*, 2006; Wu *et al.*, 2011), training tree-to-string MT models (Liu and Gildea, 2010; Aziz *et al.*, 2011), training reordering model (Xiong *et al.*, 2012) and reordering the output in the postprocessing stage (Wu and Fung, 2009). All these approaches are orthogonal to the present question of whether to train toward a semantic objective function. In fact, any of the above models could potentially benefit from the proposed approach.

2.2 Adapting SMT for formal genres to informal genres

The major challenges for machine translation in the informal genres are (1) the data demonstrates a large variety of grammar issues, such as disfluencies, incomplete sentences and misspellings; and (2) only small volumes of high-quality parallel training data are available (Mei and Kirchhoff, 2010). Rao *et al.* (2007) and Wang *et al.* (2010) proposed to remove disfluency in preprocessing stage; Bertoldi *et al.* (2010) introduced a model to recover the misspelled words before translation; Banerjee *et al.* (2011) addressed the data sparsity problem by mixing data from comparable domain into the training of both the translation model and the language model whereas He and Deng (2011) proposed to classify the training data into in-domain or out-of-domain for training two independent translation model and then combine the two models using a system combination approach. Mei and Kirchhoff (2010) incorporated document-level semantics, such as topic of the discourse, through contextual modelling. Again, all these approaches are orthogonal to our approach of incorporating semantics into SMT by tuning against a semantic objective function and any of the

above models could potentially benefit from tuning with semantic metrics.

2.3 MT evaluation metrics

Lo *et al.* (2013) showed that tuning against BLEU (Papineni *et al.*, 2002) or TER (Snover *et al.*, 2006) does not sufficiently drive SMT into making decisions to produce adequate translations. Other similar n-gram based or edit distance based metrics, such as NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006) and WER (Nießen *et al.*, 2000) also suffer from the same problem of failing to adequately reflect translation utility and correctly bias translation model for producing adequate translation.

On the other hand, no work has been done towards tuning SMT systems against more linguistically motivated MT evaluation metrics because of expensive run time costs. ULC (Giménez and Màrquez, 2007, 2008) is an aggregated metric that incorporates a large set of linguistic features, including several semantic 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). Lambert *et al.* (2006) tuned on QUEEN, a simplified version of ULC, that discards the semantic features of ULC and bases on pure lexical similarity. 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 a semantic objective function. TINE (Rios *et al.*, 2011) is a recall-oriented metric which aims to preserve the basic event structure but only performs comparably to BLEU and worse than METEOR on correlation with human adequacy judgment.

In contrast, Lo *et al.* (2013) show that a MT system tuned against MEANT (Lo *et al.*, 2012) produces more adequate translations in formal news genre as evaluated both quantitatively and qualitatively. Precisely, MEANT is computed as follows:

1. Apply an automatic shallow semantic parser on both the references 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) Compute the lexical/phrasal similarity scores to 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/phrasal similarity.

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

$$
\begin{aligned}
M_{i,j} &\equiv \text{total \# ARG j of aligned frame i in MT} \\
R_{i,j} &\equiv \text{total \# ARG j of aligned frame i in REF} \\
S_{i,\text{pred}} &\equiv \text{similarity of predicate in aligned frame i} \\
S_{i,j} &\equiv \text{similarity of ARG j in aligned frame i} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j} \\
m_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}} \\
r_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}} \\
\text{precision} &= \frac{\sum_i m_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j M_{i,j}}}{\sum_i m_i} \\
\text{recall} &= \frac{\sum_i r_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j R_{i,j}}}{\sum_i r_i}
\end{aligned}
$$

where m_{i} and r_{i} are the weights for frame i that estimate the degree of contribution of the frame to the overall meaning of the sentence in the MT/REF respectively. $M_{\text{i,j}}$ 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/phrasal similarities of the predicates and role fillers of the arguments of type j between the MT and REF. w_{pred} and w_j are the weights of the the predicates and role fillers of the arguments of type j in the MT and REF. There are 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b) and they are determined by optimizing the correlation with human adequacy judgments using grid search (Lo and Wu, 2011a).

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 in the informal genres.

Aiming at improving SMT adequacy of informal genres, we set up two experiments on public speech TED talk data and web forum data. The TED talk MT system is trained on the IWSLT2012 Chinese-English parallel TED talk training data consisting of over 130k sentences pairs. The development and

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

TED talk	BLEU ↑	NIST ↑	METEOR no_syn ↑	METEOR ↑	WER ↓	CDER ↓	TER ↓	MEANT ↑
BLEU-tuned	12.09	4.36	38.14	41.28	83.87	68.55	80.83	22.70
TER-tuned	9.63	3.67	32.75	35.19	74.00	59.24	72.31	20.41
MEANT-tuned	**11.24**	**4.22**	**38.57**	**41.96**	**80.97**	**66.21**	**78.10**	**22.74**

Table 2: Translation quality of MT system tuned against MEANT, BLEU and TER on web forum 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	17.11
TER-tuned	6.94	2.21	28.55	30.85	76.15	57.96	74.73	15.39
MEANT-tuned	**7.92**	**3.11**	**30.40**	**33.08**	**77.32**	**61.01**	**74.64**	**17.27**

test sets of the TED talk data consist of 934 and 1664 sentences respectively with one reference. The web forum MT system is trained on a large collection of Chinese-English LDC newswire data with a small portion of web forum data released under DARPA GALE and BOLT projects. The development and test sets were a held-out subset of the DARPA BOLT phase 1 web forum data which consist of 2,000 sentences and 1,697 sentences respectively with one reference. We experimented two different MT systems: a Moses phrase based MT system trained on the TED talk data and a Moses hierarchical MT system trained on the BOLT web forum data. The TED talk Moses phrase based system leverage a 6-gram language model trained on the English side of the parallel training data while the web forum Moses hierarchical system uses a 5-gram language model trained on a large volume of mixed newswire and web forum data.

We use ZMERT (Zaidan, 2009) to tune the MT system because it is a highly competitive, robust, and reliable implementation of MERT that is also fully configurable and extensible for incorporating new evaluation metrics. In this experiment, we use a MEANT 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 because it normalizes the phrase length better than geometric mean.

4 Results

Of course, tuning against any metric would maximize the performance of the SMT system on that particular metric; it would be overfitting. In the following, we avoid comparing on metrics too similar to the one that the system was tuned on. This is because Cer *et al.* (2010) showed that tuning on METEOR, TER and their variations would do well on metrics similar to what they were tuned on but perform particularly poorly on the other metrics. Therefore, it is less meaningful to evaluate a system on metrics similar to what they were tuned on.

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. In addition, we believe a good translation is one from which the reader could successfully understand at least the meaning of the source sentence, instead of just being fluent in the target language. With these as our objectives, we present the results of comparing MEANT-tuned systems against the baseline as evaluated on commonly used automatic metrics and human judgement of adequacy.

4.1 Cross-evaluation using automatic metrics

Tables 1 and 2 show that MEANT-tuned systems achieve the best scores across all other metrics on both TED talk and web forum data, when avoiding the comparison on metrics too similar to the one that the system was tuned on (the cells shaded in grey in the table). METEOR is grouped with BLEU and NIST because they are all n-gram based MT evaluation metrics. Our results indicate that MEANT-tuned system maintains a balance between lexical choices and word order as it performs well on n-gram based metrics that reward correct lexical choices 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.

Table 3: No. of sentences ranked the most adequate by human evaluators for each system in the web forum experiment.

	Eval 1	Eval 2
BLEU-tuned (B)	47	42
TER-tuned (T)	28	23
MEANT-tuned (M)	59	68
B=T	0	0
M=B	8	9
M=T	4	4
M=B=T	4	4

4.2 Human subjective evaluation

In line with our original objective of biasing SMT systems toward producing adequate translations, we conduct a manual evaluation to judge the translation utility of outputs produced by MEANT-, BLEU- and TER-tuned systems in the web forum experiment. Following the manual evaluation protocol of Lambert *et al.* (2006), we randomly sampled 150 sentences from the MT output of the three systems in web forum domain. The output of our system and the two baselines along with the input sentence and the reference translation was presented to human evaluators. Two evaluators were instructed to choose the most adequate translation from the three MT output. The inter-evaluator agreement was 70%.

Table 3 indicates that output of the MEANT-tuned system is ranked adequate more frequently compared to BLEU- and TER-tuned baselines. We performed the right-tailed two proportion significance test on human evaluation of the SMT system outputs for both the genres. The MEANT-tuned system generates more adequate translations on web forum data than the TER-tuned system at the 99% significance level. The MEANT-tuned system is ranked more adequate than the BLEU-tuned system at the 95% significance level. 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.

Although one might expect an SRL dependent metric such as MEANT to perform poorly on the domain of informal text, 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 toward adequate output on informal text domains despite the shortcomings of automatic shallow semantic parsing.

5 Error analysis

To better understand why the MEANT-tuned MT system still able to outperform the BLEU- and the TER-tuned system despite reports pointing out the deficiency of automatic semantic parser on informal text data, we conduct a thorough error analysis. We specifically look into cases when the automatic shallow semantic parser fails to construct a parse from the sentence. Besides analysing the development and the test set of the two previous mentioned experiments, we also look for the same phenomenon in the MetricsMaTr 2008 broadcast news data set to ensure a more reliable result from the error analysis.

Table 4 shows that over 14% of the sentences in the TED talk demonstrate no semantic parse and on average over 8% of the sentences in the web forum data set has no semantic parse. The failure of the automatic semantic parser to provide any parse for MEANT to score the sentence would result in a zero MEANT score on those sentences. We further investigate into the cases when the automatic semantic parser fails to identify any semantic frames in the sentences so as to understand how to incorporate semantic information in those cases.

5.1 Failure to label the "be" semantic frames

Surprisingly, Table 5 shows that the failure of automatic semantic parsing on identifying semantic frames did not result from the ungrammatical sentences in the informal data. Instead, the major source of errors in automatic shallow semantic parsing of informal genres is failure to identify the semantic frame for copula or existential senses of "be" in perfectly grammatical sentences. The following is the Propbank (Palmer *et al.*, 2005) framesets definition of the predicate "*be*":

> Roleset *be.01*: *copula*
> Roles: ARG1-*topic*, ARG2-*comment*
>
> Roleset *be.02*: *existential*
> Roles: ARG1-*thing that is*
>
> Roleset *be.03*: *auxiliary*
> Roles: **do not tag**

The following are the examples from the TED talk and the web forum data showing the usage of the three senses of "be":

Table 4: Number of sentences with no automatic semantic parsing output in each data set

dataset	genre	#sentences	#no semantic parse	%no semantic parse
TED-dev	public talk	934	138	14.78%
TED-test	public talk	1664	237	14.24%
BOLT P1-dev	forum	2000	229	11.45%
BOLT P1-test	forum	1697	100	5.89%
MetricsMaTr 08	broadcast news	221	9	4.07%

Table 5: Detailed breakdown of the sentences with no semantic frame identified by the automatic semantic parser. (#"be" is the number of sentences that has at least one grammatical and valid semantic frame of the copula or existential sense of "be"; #no verb ($\leq$ 10) and #no verb ($>$ 10) are the number of sentences that has no verb in the sentence with the sentence length is "less than or equal to 10" or "greater than 10" respectively; #other is the number of sentences that do not fall into any of the previous categories.)

dataset	genre	#no parse	#"be"	#no verb ($\leq$10)	#no verb ($>$ 10)	#others
TED-dev	public talk	138	110	20	3	5
TED-test	public talk	237	191	38	3	5
BOLT P1-dev	forum	229	166	56	6	1
BOLT P1-test	forum	100	81	4	5	10
MetricsMaTr 08	broadcast news	9	9	0	0	0

Example 5.1 (copula)
A language is a flash of the human spirit .

Example 5.2 (existential)
There is no feed .

Example 5.3 (auxiliary) [ARG0 *The sun*] *is* [PRED *rising*] .

In Example 5.1 and Example 5.2, the automatic semantic parser fails to identify the verb "be" as the predicate and thus fails to construct the corresponding semantic frame. In Example 5.3, the predicate "rising" is correctly identified and the auxiliary sense of "be" is not tagged following the Propbank frameset definition.

We believe that the failure is due to the differences in language usage between formal and informal genres. In formal text genres (i.e. which the automatic semantic parser is trained on), "be" is always used as auxiliary verb together with the present or past participle to realize different tense or voice in grammar. Since Propbank declares that the auxiliary sense of "be" should not be labeled, the automatic semantic parser is biased heavily not to label "be" as the predicate. However, as we can see in Table 5, up to 11% of the sentences in informal genres are having the copula or the existential sense of "be" as the predicate.

Apart from the problem of "be", another common situation for automatic parser fails to identify semantic frames in informal text data is that there is no predicate verb at all in the sentence. We breakdown the statistics of sentences with no semantic frame according to the sentence length and find that most of these sentences are short phrases with less than 10 words. For these cases, from experiments on similar phenomena, we observed that using the phrasal similarity function for the role fillers in MEANT was still better than BLEU or other n-gram based metrics for evaluating translation adequacy at sentence-level, because only very rarely do these sentences have sufficient n-gram counts for the metrics to accurately differentiate translation quality. The following are the examples of the no-verb sentences from the TED talk and the web forum data:

Example 5.4 (no verb, sentence length$\leq$10)
2011 Summer Davos Forum

Example 5.5 (no verb, sentence length$>$10)
A photo of Wang Lei , counselor for the Arts Department 2008 , Henan University

5.2 Reconstructing the frames of "be"

The results of the error analysis have lead to another interesting research question: is it possible to further improve the performance of the MEANT-tuned SMT system by reconstruct the missing semantic frames of "be" correctly? We manually reconstruct the missing semantic frames of "be" in the MetricsMaTr 2008 data set. We compute the MEANT scores of the three system outputs in

the MetricsMaTr data set and correlate the scores against the adequacy judgments as provided in the MetricsMaTr 2008 data set.

With the reconstructed semantic frames of "be", the correlation of MEANT with human adequacy judgment on MetricsMaTr 2008 broadcast news data set improves from 14.67 to 16.00. These results suggest that correctly reconstructing the missing semantic frames of "be" would possibly serve as a future basis for further improvements the translation quality for informal text.

6 Conclusion

We have presented the first ever results to demonstrate that even for informal language genres, tuning an SMT system against MEANT robustly produces more adequate translation than tuning against BLEU or TER, as measured across a battery of commonly used automated metrics, as well as via human subjective evaluation. Surprisingly, tuning against MEANT succeeds in producing adequate output significantly more frequently even though it depends on automatic semantic parses which are notoriously hard to find in these informal genres. We argue that by rewarding preservation of even a portion of the meaning of the translations as captured by semantic frames during the training process, the gains from constraining the SMT system to make more accurate lexical choices and reordering outweigh the losses from fragile automatic SRL on informal language.

We believe that tuning on MEANT will prove equally useful for SMT systems based on any architecture, particularly where the model does not otherwise incorporate semantic information to improve the adequacy of the translations produced. Our encouraging results even on informal language suggest that using MEANT-like objective functions to tune SMT would drive sustainable development of MT toward higher utility. Future work of this line of research includes using more stable and efficient optimization techniques, such as MIRA or PRO, to tune MT systems against MEANT; and tuning MT systems for other language pairs against MEANT.

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.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. *Proceedings of Machine Translation Summit XIII, Xiamen, China*, pages 285–292, 2011.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419, 2010.

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.

Daniel Cer, Cristopher D. Manning, and Daniel Jurafsky. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT-10)*, pages 555–563, 2010.

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.

Benoit Favre, Bernd Bohnet, and D Hakkani-Tür. Evaluation of semantic role labeling and dependency parsing of automatic speech recognition output. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5342–5345. IEEE, 2010.

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.

Xiaodong He and Li Deng. Robust speech translation by domain adaptation. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, volume 201, page l, 2011.

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.

Xiaohua Liu, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Zhongyang Xiong, and Changning Huang. Semantic role labeling for news tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 698–706, 2010.

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, 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.

Yang Mei and Katrin Kirchhoff. Contextual modeling for meeting translation using unsupervised word sense disambiguation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pages 1227–1235, 2010.

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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: an Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

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.

Sharath Rao, Ian Lane, and Tanja Schultz. Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. *Training*, 6370(46300):6–50, 2007.

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.

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.

Wen Wang, Gokhan Tur, Jing Zheng, and Necip Fazil Ayan. Automatic disfluency removal for improving spoken language translation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5214–5217. IEEE, 2010.

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.