Can Semantic Role Labeling Improve SMT?

Dekai Wu[1] **Pascale Fung**[2]

Human Language Technology Center
HKUST
[1]Department of Computer Science and Engineering
[2]Department of Electronic and Computer Engineering
University of Science and Technology, Clear Water Bay, Hong Kong
dekai@cs.ust.hk pascale@ee.ust.hk

Abstract

We present a series of empirical studies aimed at illuminating more precisely the likely contribution of semantic roles in improving statistical machine translation accuracy. The experiments reported study several aspects key to success: (1) the frequencies of types of SMT errors where semantic parsing and role labeling could help, and (2) if and where semantic roles offer more accurate guidance to SMT than merely syntactic annotation, and (3) the potential quantitative impact of realistic semantic role guidance to SMT systems, in terms of BLEU and METEOR scores.

1 Introduction

In this investigative paper, we present a new set of empirical studies aimed at illuminating more precisely the likely contribution of semantic parsing and role labeling toward improving statistical machine translation accuracy.

The most glaring errors made by statistical machine translation systems continue to be those resulting in confusion of semantic roles. These sorts of translation errors often result in serious misunderstandings of the essential meaning of the source utterances — who did what to whom, for whom or what, how, where, when, and why.

It has been widely observed that the negative impacts of such errors on the utility of the translation are inadequately reflected by evaluation metrics based on lexical criteria. The accuracy of translation lexical choice has reached increasingly satisfactory levels—at least for largely literal genres such as newswire — which helps boost lexically oriented scores such as BLEU (Papineni *et al.*, 2002) or METEOR (Banerjee and Lavie, 2005) despite serious role confusion errors in the translations.

It has also often been noted that precision-oriented metrics such as BLEU tend to reward fluency more than adequacy (in particular, BLEU's length penalty is only an indirect and weak indicator of adequacy). Today's SMT systems produce translations that often contain significant role confusion errors but nevertheless read quite fluently.

Thus, while recent years have seen continued improvement in the accuracy of statistical machine translation systems as measured by such lexically based metrics, this underestimates the effect of the persistent errors of role confusion upon the actual translation utility.

This situation leads us to consider the potential application of shallow semantic parsing and semantic role labeling models to SMT, in ways that might reduce role confusion errors in the translation output. Within the lexical semantics community, increasingly sophisticated models for shallow semantic parsing are being developed. Such semantic parsers, which automatically label the predicates and arguments (roles) of the various semantic frames in a sentence, could automatically identify inconsistent semantic frame and role mappings between the input source sentences and their output translations. This approach is supported by the results of Fung *et al.* (2006), which reported that (for the Chinese-English language pair) approximately 84% of semantic role mappings remained consistent cross-lingually across sentence translations.

We approach this promise with caution, however, given the painful lessons learned through the historical difficulty of making syntactic and semantic models contribute to improving SMT accuracy. The past decade has at last seen increasing amounts of evidence that SMT accuracy can indeed be improved via tree-structured and syntactic models (e.g., Wu (1997); Chiang and Wu (2008); Wu and Chiang (2009)) despite numerous disappoint-

© 2009 European Association for Machine Translation.

Proceedings of the 13th Annual Conference of the EAMT, pages 218–225,
Barcelona, May 2009

ing attempts Och *et al.* (2004). More recently, lexical semantics models for word sense disambiguation have also finally been successfully applied to increasing SMT accuracy (e.g., Carpuat and Wu (2007), Chan *et al.* (2007); Giménez and Màrquez (2007a)) again after surprising initial failures (e.g., Carpuat and Wu (2005)). In both the syntactic and semantic cases, improving SMT accuracy ultimately required making major adaptations to the original linguistic models. We can reasonably expect it to be at least as difficult to successfully adapt the even more complex types of lexical semantics modeling from semantic parsing and role labeling.

Avoiding the many potential blind alleys calls for careful analysis and evaluation of (1) the frequencies of types of SMT errors where semantic parsing and role labeling could help, (2) if and when semantic roles offer more accurate guidance to SMT than merely syntactic annotation, and (3) the potential quantitative impact of realistic semantic role guidance to SMT systems, at least in terms of scores such as BLEU and METEOR.

In this paper, we present a series of four experiments designed to address each of these questions, using Chinese-English parallel resources, a typical representative SMT system based on Moses, and shallow semantic parsers for both English and Chinese.

2 Related work

While this is a new avenue of inquiry, the background relevant to the experiments described here includes (1) a broad body of work on shallow semantic parsing and semantic role labeling, the majority of which has been performed on English, (2) a relatively small body of work specific to semantic parsing and semantic role labeling of Chinese, and (3) a proposal to measure semantic role overlap as one of the key factors in new MT evaluation metrics.

2.1 Shallow semantic parsing

Semantic parsers analyze a sentence with the aim of identifying the "who did what to whom, for whom or what, how, where, when, and why." Shallow semantic parsing extracts the predicate-argument structure of verbs in a sentence based on the syntactic tree of that sentence. For example, the predicate argument structure of the verb *hold* in Figure 1 specifies a "holding" relation between *both sides* (who) and *meeting* (what) *on Sunday*

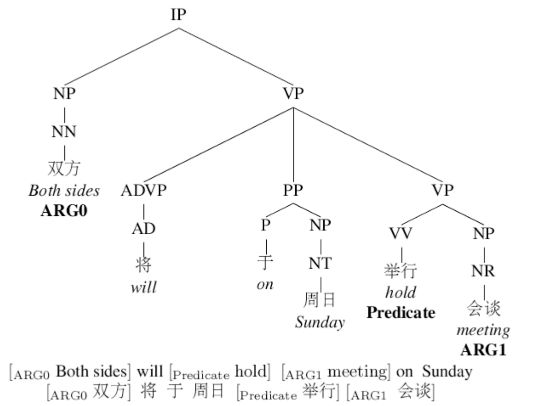

Figure 1: Chinese shallow semantic parsing example.

(when). For a sentence with multiple verbs, there can be multiple predicate argument structures.

Shallow semantic parsing systems are mostly based on classifiers that learn from a manually annotated semantic corpus (Gildea and Jurafsky (2002), Pradhan *et al.* (2005)). Following the publication of the Proposional Bank (PropBank) (Palmer *et al.*, 2005) first in English, then in Chinese, it has been possible to train these classifiers to perform semantic analysis on news wire type of texts.

2.2 Chinese shallow semantic parsing

Systems that perform shallow semantic parsing on Chinese texts are likewise based on classifiers and trained on the Chinese PropBank and the bilingual Chinese-English Parallel PropBank (Sun and Jurafsky (2004), Xue (2006), Fung *et al.* (2006)). It is interesting to note that, despite the very different characteristics of Chinese verbs (Xue and Palmer, 2005) from those in English, the core algorithm of a shallow semantic parser remains the same. As was found to be the case in English, SVM classifiers have been found to outperform maximum entropy classifiers for this task (Fung *et al.*, 2006). The primary difference lies in the feature set chosen to represent semantic information.

In experiments carried out on PropBank data using gold standard syntactic parse trees, extended syntactic features such as Path Trigram and Path Abbreviations were found to have the highest contribution to system performance (Fung *et al.*, 2006). Another feature, Verb Cluster, was also found to be most useful by Xue and Palmer (2005).

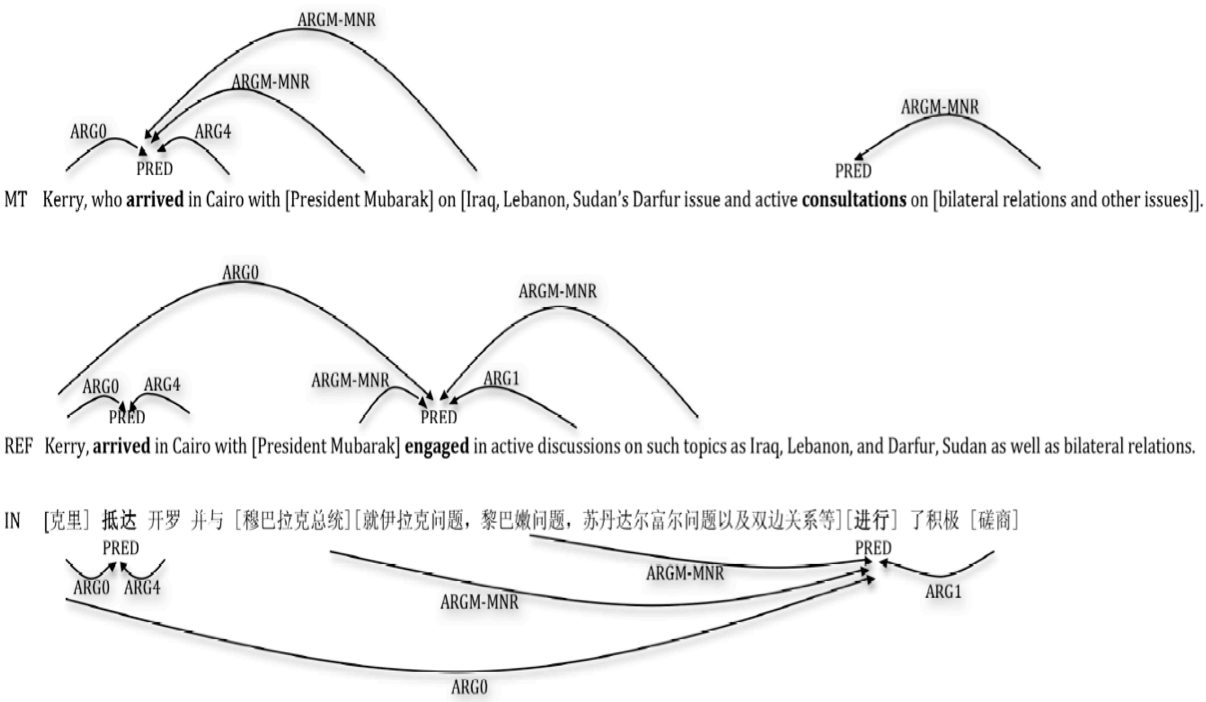

Figure 2: Example of semantic frames in Chinese input and English MT output.

2.3 MT evaluation metrics based on semantic role overlap

Giménez and Màrquez (2007b) and Giménez and Màrquez (2008) introduced and refined a set of new MT evaluation metrics employing rich assortments of features reflecting various kinds of similarity at lexical, shallow syntactic, deep syntactic, shallow semantic, and deep semantic levels.

Under a number of scenarios—particularly the out-of-domain scenarios — measuring the overlap of shallow semantic roles between the source and target language sentence pairs contributes to improved correlation with human judgment of translation quality. Unsurprisingly, measuring the overlap of manually annotated deep semantic relations contributes even more in some scenarios. However, given the state of automatic semantic parsing technology, realistically we are today still much closer to being able to incorporate automatic shallow semantic parsing into working SMT systems, and thus we focus on shallow semantic parsing and semantic role labeling for the present.

3 Semantic frames in SMT output

The first of the experiments aims to provide a more concrete understanding of one of the key questions as to the role of semantic parsing in SMT: how well do typical current SMT systems already perform on semantic frames?

The annotated example in Figure 2 shows, from bottom to top, (IN) a fragment of a typical Chinese input source sentence that is drawn from newswire text, (REF) the corresponding fragment from its English reference sentence, and (MT) the corresponding fragment of the output sentence from a state-of-the-art SMT system.

A relevant subset of the semantic roles and predicates has been annotated in these fragments. In the Chinese input and its corresponding English reference, there are two main verbs marked PRED. The first, (*arrived*), has two arguments: one in an ARG0 agent role, (*Kerry*); and another in an ARG4 destination role, (*Cairo*). The second verb, (*engaged*), has four arguments: one in an ARG0 agent role, again (*Kerry*); one in an ARG1 role, (*discussions*); and two others in ARGM-MNR manner roles, (*with Mubarak*) and (*on topics*).[1]

In contrast, in the SMT translation output, a very different set of predicates and arguments is seen. While the PRED *arrived* still has the same correct ARG0 *Kerry* and ARG4 *Cairo*, now the ARGM-MNR manner role with *President Mubarak* is incorrectly modifying the *arrived*, instead of an engaged predicate. In fact, the *en-*

[1]Minor variations on the role labeling in these examples are possible, but not central to the present point.

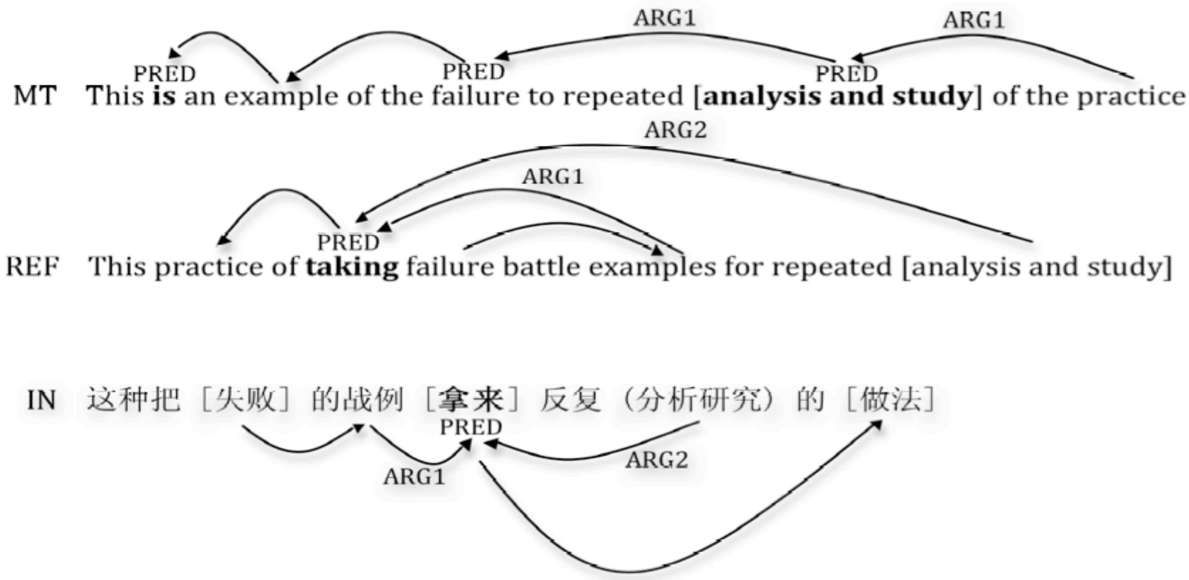

Figure 3: Example of semantic frames in Chinese input and English MT output.

gaged predicate has erroneously been completely dropped by the SMT system, so there is no verb to which the arguments of *engaged* can be attached.

Figure 3 shows another typical example. Again, PRED marks the main verb in the Chinese input source fragment and its corresponding English reference, (*taking*). It has two arguments: an ARG1 (*battle examples*) and an ARG2 (*analysis and study*).

The SMT translation output, however, not only lacks the main verb, but includes many incorrect predicates and roles. Such spurious predicate-argument structures are clearly seriously detrimental to even cooperative readers straining to guess the meaning of the original Chinese.

3.1 Experimental setup

To assess the above sorts of phenomena quantitatively, we designed an experiment making use of 745 bi-sentences extracted from the Parallel PropBank with gold standard annotations of both syntactic and semantic roles.

We use the Chinese sentences as system input and their corresponding English translations as the reference translations. We use the open source statistical machine translation decoder Moses (**?**) for the experiments, translating the PropBank Chinese sentences into English with the same model trained for our participation in the IWSLT 2007 evaluation campaign (Shen *et al.*, 2007).The English translations generated by the decoder are the system output. Based on the system input and the reference

Table 1: Accuracy of predicate-argument structure in Chinese-English SMT output for data set A.

P-A Structure	Precision	Recall	F-measure
Predicate	0.98	0.57	0.72
ARG0	0.74	0.38	0.50
ARG1	0.73	0.41	0.53
ARG2	0.82	0.32	0.46
ARG3	1.00	0.67	0.80
ARG4	1.00	0.33	0.51
All ARGs	0.74	0.39	0.51

translation, we intend to investigate whether the predicate verbs are correctly translated and their predicate-argument structures preserved in the system output.

We first randomly select 50 bi-sentences, without any constraint on the translation accuracy of the predicate verbs, to form the first observation data set (data set A).

3.2 Experimental results

Human evaluation of these results show that, for all 138 predicate verbs in the system input (Chinese sentences), only 79 (around 57%) of them are correctly translated in the system output; and given such correctly translated predicate verbs, the translation of their semantic arguments can only achieve around 51% overall F-measure. The detailed results are shown in Table 1.

Table 2: Accuracy of predicate-argument structure in Chinese-English SMT output for data set B.

P-A Structure	Precision	Recall	F-measure
Predicate	1.00	1.00	1.00
ARG0	0.83	0.66	0.74
ARG1	0.84	0.78	0.81
ARG2	0.80	0.78	0.38
ARG3	0.00	0.00	N/A
ARG4	0.50	1.00	0.66
All ARGs	0.84	0.68	0.75

43% of the Chinese predicate verbs are either not translated at all into English or are translated into a different part-of-speech category such as nouns or adjectives. As shown in Figure 4, the predicate verb 位/*located* in the input Chinese sentence is not translated in the system output

4 Semantic roles in SMT output

In the previous experiment, the semantic role accuracy in output translations was negatively affected by errors in identifying the central verb in the first place — as we have seen in both introductory examples of Section 3 as well as the example of Figure 1. Without the verb, properly identifying the arguments becomes meaningless. It is therefore worth asking a secondary version of the question: *providing the verb is correctly translated*, then how well do typical current SMT systems perform on semantic roles?

4.1 Experimental setup

Since nearly half of the predicate verbs in the system input are not translated or wrongly translated in the system output in the previous experiment, we construct another data set (data set B) by randomly selecting 50 bi-sentences under an additional constraint that all predicate verbs are correctly translated. We carry out the same analysis on data set B and the result is shown in Table 2.

4.2 Experimental results

For data set B, the overall F-measure of the translation of the semantic arguments is about 75%, which is 24 points higher than that in data set A.

In this data set B, we also find that some of the semantic roles are missing in the system output.

A common type of translation error occurs when a group of words that together have a single semantic role in the source language (Chinese) are split into separate groups in the translation (English) often in the wrong word order. In the example of Figure 5, the phrase 其所有资产的偿债率 in the input is translated into two separate phrases in the output: *its debt rate* and *of the assets*, creating different semantic relationships compared to the original semantic role of the source phrase. Finally, even though all the words in the arguments of a certain predicate verb are correctly translated into English, their semantic roles are found to be confusing in the translation leading to ambiguity in the interpretation of the translated sentences. As shown in the example of Figure 6, although words in both ARG0 and ARG1 are correctly translated into English, we still cannot understand the final translated sentence because the semantic roles of these two phrases are confused. We cannot tell which semantic roles *Myanmar and Thailand's government* and *the two countries border trade agreements* are supposed to play. This confusion arises from the incorrect position of the predicate verb *signed*.

As we can see, the types of translation errors shown in the examples of Figures 3 to 5 lead to ambiguity in the final understanding of the translation even though the system output still reads fluently. This is caused by the fact that current n-gram based SMT systems are not designed to take semantic roles into consideration.

5 Semantic vs. syntactic roles

The third experiment aims to answer another key question: if we favor semantic role consistency across both the source input sentence and the output translation, would this outperform merely favoring syntactic role consistency across the bisentence? In other words, does incorporating semantic role analysis contribute anything beyond the current work on syntactic SMT models?

5.1 Experimental setup

To address this question, we perform a different analysis of the previously described set of 745 bilingual sentence pairs with manually annotated syntactic and semantic roles from the Parallel PropBank.

The syntactic roles are manually annotated according to Treebank guidelines. Whereas the Chinese sentences are annotated with both "subject" and "object" syntactic roles, their English counterparts are only annotated with "subject" roles with-

IN [ARG0 上述 开发区] 基本 [PRED 位] 于 [ARG1 福建 经济 最为 活跃 的 东南部 地区] 。

REF The above-mentioned development zones are basically [PRED located] in the southeastern area of Fujian whose economy is the most active.

MT The basic development zones in the southeastern region of Fujian ' s economy is the most active.

Figure 4: Example of semantic frames in Chinese input and English MT output.

IN 去年四月， ＣＲ 公司开始了其破产程序，[ARG0 其所有资产的偿债率] 仅 [PRED 为] [ARG1 百分之五] 。

REF In April of last year, the CR Company began bankruptcy procedures and [ARG0 the debt compensation rate of all its assets] [PRED was] only [ARG1 5 %].

MT In April of last year, the company began bankruptcy procedures, all of its debt rate [PRED was] only [ARG1 five percent] [ARG0 of the assets] of the CR.

Figure 5: Example of semantic frames in Chinese input and English MT output.

Table 3: Syntactic role mapping in Chinese (ZH) to English (EN) translations.

Syntactic role mapping	Freq	Pct
ZH subject ↔ EN subject	514	84.26%
ZH subject ↔ EN NP	44	7.21%
ZH subject ↔ EN PP	31	5.08%
ZH subject ↔ EN S	15	2.46%
ZH subject ↔ EN other	6	0.98%

out the “object” roles.

Furthermore, we manually align the predicate argument structures across the bi-sentences for our experiment.

The experiment is done as follows:

1. We first extract all predicate argument structure mappings from the manually annotated and structurally aligned corpus. We compute the statistics of direct semantic role mappings (ARG*i* to ARG*i*) based on the translation.

2. From the output of step 1, we further look at the syntactic roles associated with each bilingual argument mapping. We use the semantic role boundaries from the annotated corpus to find the syntactic roles.

3. The corresponding Chinese/ English syntactic roles are then constructed as syntactic role mappings.

5.2 Experimental results

Given all the direct semantic role mappings from Chinese to English, their corresponding subject syntactic role mappings are listed below in Table 3. We can see that only 84.26% of direct semantic role mappings result from direct syntactic role projections. More than 15% of the subjects are not translated into subjects, even though their semantic roles are preserved across language.

This result shows that semantic roles enforce cross-lingual translation patterns more correctly than syntax. Whereas syntactic roles vary for each language, semantic roles that convey the meaning of a sentence are translingual.

6 Improving SMT with semantic frames

In the fourth experiment, we aim to assess the potential quantitative impact of realistic semantic role guidance to SMT systems, in terms of BLEU and METEOR scores. This is done by simulating the effect of enforcing consistency between the semantic predicates and arguments across both the input source sentence and the translation output.

6.1 Experimental setup

For this experiment, we return to data set B, as described in Section 4.1. For each sentence, two types of semantic parse based corrections are permitted to the output translation.

First, the constituent phrases corresponding to either the predicates or the arguments for any la-

IN [ARG0 缅甸 和 泰国 政府] 今天 下午 在 此间 [PRED 签订] 了 [ARG1 两 国 边境 贸易 协定]。

REF This afternoon [ARG0 the Myanmaran and Thai governments] [PRED signed] [ARG1 an agreement on border trade between their two countries] here.

MT [ARG? Myanmar and Thailand ' s government] of [ARG? the two countries border trade agreements] [PRED signed] here this afternoon.

Figure 6: Example of semantic frames in Chinese input and English MT output.

IN 加工贸易在广东外经贸发展中占有举足轻重的地位，同时也是粤港澳台经贸合作的重要内容。

REF The processing trade occupies a crucial position in the development of foreign economy and trade in Guangdong and at the same time is important content in the economic and trade cooperation between Guangdong , Hong Kong , Macao and Taiwan.

MT In the processing trade in Guangdong ' s foreign trade and economic development in Guangdong , Hong Kong , Macao , Taiwan occupies a decisive position at the same time , it is an important content of the economic and trade cooperation.

RE-ORDERED In the processing trade occupies a decisive position in Guangdong ' s foreign trade and economic development at the same time , it is an important content of the economic and trade cooperation in Guangdong , Hong Kong, Macao , Taiwan.

Figure 7: Example of semantic frames in Chinese input and English MT output.

Table 4: SMT performance improvement with semantic predicate and role consistency constraints.

Metric	Baseline translation	Enforcing consistent semantic parses
BLEU	34.76	36.62
METEOR	63.5	65.9

beled semantic role are permitted to be re-ordered such that a semantic parse of the re-ordered translation consistently matches the role label on the corresponding phrase in the input source sentence.

Second, if the translation of a predicate in the input source sentence is missing in the output translation, then a translation of that predicate may be added to the output translation such that, again, a semantic parse of the translation consistently associates it with the corresponding arguments for that predicate.

6.2 Experimental results

The results, as shown in Table 4, show that favoring semantic frame and role consistency across the source input sentence and the output translation improves BLEU and METEOR scores. The accuracy improves on the order of two points, for both metrics.

The example of Figure 7 shows how two of the constituent phrases are re-ordered.

It is worth noting that both BLEU and METEOR are still n-gram based metrics, which are of limited accuracy at evaluating fine-grained semantic distinctions in the translations. We suspect that the enhancement in translation quality would be even more obvious under utility-based MT evaluation strategies; this is one main direction for future research.

7 Conclusion

We have presented a series of experimental studies that illuminate more precisely the likely contribution of semantic roles in improving statistical machine translation accuracy. The experiments reported studied several aspects key to success: (1) the frequencies of types of SMT errors where semantic parsing and role labeling could help, and (2) if and where semantic roles offer more accurate guidance to SMT than merely syntactic annotation, and (3) the potential quantitative impact of realistic semantic role guidance to SMT systems, in terms of BLEU and METEOR scores. All sets of results support the utility of shallow semantic parsing and semantic role labeling for improving certain limited but important aspects of SMT accuracy.

Our studies have focused on Chinese and English. Chinese and English are of course semantically very different, arising from their completely unrelated origins in the Sino-Tibetan and European language families. The effect is seen in the fact that state-of-the-art machine translation accuracy remains low for Chinese-English, even though intensive research on other "difficult" language pairs such as Arabic-English began far more recently. We conjecture that similar or better results to those reported in this paper would hold for most other language pairs, especially where there are closer correspondences in the semantic frame inventory of the two languages.

Acknowledgments We would like to thank Zhaojun Wu, Yongsheng Yang, Chi-kiu Lo, Huaqing Luo, Yihai Shen, and Marine Carpuat for their assistance. This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants GRF621008, GRF612806, DAG03/04.EG09, RGC6256/00E, and RGC6083/99E. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 387–394, Ann Arbor, Jun 2005.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP– CoNLL 2007)*, pages 61– 72, Prague, Jun 2007.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL–07)*, Prague, June 2007.

David Chiang and Dekai Wu, editors. *Proceedings of SSST–2, Second Workshop on Syntax and Structure in Statistical Translation, at ACL– 08:HLT*. Association for Computational Linguistics, Columbus, OH, Jun 2008.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. Automatic learning of chinese–english semantic structure mapping. In *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*, Aruba, Dec 2006.

Daniel Gildea and Dan Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):256–288, 2002.

Jesús Giménez and Lluís Màrquez. Context–aware discriminative phrase selection for statistical machine translations. In *Workshop on Statistical Machine Translation*, Prague, Jun 2007.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous mt systems. In *Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Jun 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic mt evaluation. In *3rd ACL Workshop on Statistical Machine Translation*, pages 195– 198, Columbus, OH, Jun 2008.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison– Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL–2007 Demo and Poster Sessions*, pages 177–180, Prague, Jun 2007.

Franz Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. A smorgasbord of features for statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL–2004)*, Boston, May 2004.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, Mar 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei–Jing Zhu. BLEU: A method for automatic evaluation of machine translations. In *40th Annual Meeting of the Association for Computational Linguistics (ACL–2002)*, pages 311–318, Philadelphia, Jul 2002.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. Support vector learning for semantic argument classification. *Machine Learning*, 60(1–3):11–39, 2005.

Yihai Shen, Chi-kiu Lo, Marine Carpuat, and Dekai Wu. HKUST statistical machine translation experiments for IWSLT 2007. In *Fourth International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 84–88, Trento, Oct 2007.

Honglin Sun and Daniel Jurafsky. Shallow semantic parsing of chinese. In *Human Language Technology Conference/ North American Chapter of the Association for Computational Linguistics (HLT/ NAACL– 2004)*, pages 249– 256, Boston, May 2004.

Dekai Wu and David Chiang, editors. *Proceedings of SSST–3, Third Workshop on Syntax and Structure in Statistical Translation, at NAACL–HLT 2009*. Association for Computational Linguistics, Boulder, CO, Jun 2009.

Dekai Wu. Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, Sep 1997.

Nianwen Xue and Martha Palmer. Automatic semantic role labeling for chinese verbs. In *19th International Joint Conference on Artificial Intelligence (IJCAI– 05)*, Edinburgh, 2005.

Nianwen Xue. Semantic role labeling of nominalized predicates in chinese. In *Human Language Technology Conference of the North American Chapter of the ACL (HLT–NAACL 2006)*, pages 431–438, New York, Jun 2006.