Semantic Roles for SMT: A Hybrid Two-Pass Model

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 results on a novel hybrid semantic SMT model that incorporates the strengths of both semantic role labeling and phrase-based statistical machine translation. The approach avoids major complexity limitations via a two-pass architecture. The first pass is performed using a conventional phrase-based SMT model. The second pass is performed by a re-ordering strategy guided by shallow semantic parsers that produce both semantic frame and role labels. Evaluation on a Wall Street Journal newswire genre test set showed the hybrid model to yield an improvement of roughly half a point in BLEU score over a strong pure phrase-based SMT baseline – to our knowledge, the first successful application of semantic role labeling to SMT.

1 Introduction

Many of the most glaring errors made by today's statistical machine translation systems are those resulting from confusion of semantic roles. Translation errors of this type frequently result in critical misunderstandings of the essential meaning of the original input language sentences – who did what to whom, for whom or what, how, where, when, and why.

Semantic role confusions are errors of adequacy rather than fluency. It has often been noted that the dominance of lexically-oriented, precision-based metrics such as BLEU (Papineni *et al.* 2002) tend to reward fluency more than adequacy. The length penalty in the BLEU metric, in particular, is only an indirect and weak indicator of adequacy. As a result, SMT work has been driven to optimize systems such that they often produce translations that contain significant role confusion errors despite reading fluently.

The present work is inspired by the question of whether we can improve translation utility via a strategy of favoring semantic adequacy slightly more – possibly at the expense of slight degradations in lexical fluency.

Shallow semantic parsing models have attained increasing levels of accuracy in recent years (Gildea and Jurafsky 2000; Sun and Jurafsky 2004; Pradhan *et al.* 2004, 2005; Pradhan 2005; Fung *et al.* 2006, 2007; Giménez and Màrquez 2007a, 2008). Such models, which identify semantic frames within input sentences by marking its predicates, and labeling their arguments with the semantic roles that they fill.

Evidence has begun to accumulate that semantic frames – predicates and semantic roles – tend to preserve consistency across translations better than syntactic roles do. This is, of course, by design; it follows from the definition of semantic roles, which are less language-dependent than syntactic roles. Across Chinese and English, for example, it has been reported that approximately 84% of semantic roles are preserved consistently (Fung *et al.* 2006). Of these, roughly 15% do *not* preserve syntactic roles consistently.

Since this directly targets the task of determining semantic correctness, we believe that the adequacy of MT output could be improved by leveraging the predictions of semantic parsers. We would like to exploit automatic semantic parsers to identify inconsistent semantic frame and role mappings between the input source sentences and their output translations.

However, we take note of the difficult experience in making syntactic and semantic models con-

Proceedings of NAACL HLT 2009: Short Papers, pages 13–16,
Boulder, Colorado, June 2009.

tribute to improving SMT accuracy. On the one hand, there is reason to be optimistic. Over the past decade, we have seen an accumulation of evidence that SMT accuracy can be improved via tree-structured and syntactic models (e.g., Wu 1997; Wu and Chiang 2009), and more recently, work from lexical semantics has also at long last been successfully applied to increasing SMT accuracy, in the form of techniques adapted from word sense disambiguation models (Chan *et al.* 2007; Giménez and Màrquez 2007b; Carpuat and Wu 2007). On the other hand, both directions saw unexpected disappointments along the way (e.g., Och *et al.* 2003; Carpuat and Wu 2005). We are therefore forewarned that it is likely to be at least as difficult to successfully adapt the even more complex types of lexical semantics modeling from semantic parsing and role labeling to the translation task.

In this paper, we present a novel hybrid model that, for the first time to our knowledge, successfully applies semantic parsing technology to the challenge of improving the quality of Chinese-English statistical machine translation. The model makes use of a typical representative SMT system based on Moses, plus shallow semantic parsers for both English and Chinese.

2 Hybrid two-pass semantic SMT

While the accuracy of shallow semantic parsers has been approaching reasonably high levels in recent years for well-studied languages like English, and to a lesser extent, Chinese, the problem of excessive computational complexity is one of the primary challenges in adapting semantic parsing technology to the translation task.

Semantic parses, by definition, are less likely than syntactic parses to obey clearly nested hierarchical composition rules. Moreover, the semantic parses are less likely to share an exactly isomorphic structure across the input and output languages, since the raison d'être of semantic parsing is to capture semantic frame and role regularities independent of syntactic variation – monolingually and cross-lingually.

This makes it difficult to incorporate semantic parsing into SMT merely by applying the sort of dynamic programming techniques found in current syntactic and tree-structured SMT models, most of which rely on being able to factor the computation into independent computations on the subtrees. In other words, the key computational obstacle is that the semantic parse of a larger string (or string pair, in the case of translation) is not in general strictly mechanically composable from the semantic parses of its smaller substrings (or substring pairs).

In fact, the lack of easy compositionality is the reason that today's most accurate shallow semantic parsers rely not primarily on compositional parsing techniques, but rather on ensembles of predictors that independently rate/rank a wide variety of factors supporting the role assignments given a broad *sentence-wide* range of context features. But while this improves semantic parsing accuracy, it poses a major obstacle for efficient tight integration into the sub-hypothesis construction and maintenance loops within SMT decoders.

To circumvent this computational obstacle, the hybrid two-pass model defers application of the non-compositional semantic parsing information until a second error-correcting pass. This imposes a division of labor between the two passes.

---

1. Apply a semantic parser for the *input* language to the input source sentence.
2. Apply a semantic parser for the *output* language to the baseline translation that was output by the first pass. Note: this also produces a shallow syntactic parse as a byproduct.
3. If the semantic frames (target predicates and their associated semantic roles) are all consistent between the input and output sentences, and are aligned to each other by the phrase alignments from the first pass, then finish immediately and output the baseline translation.
4. Segment the baseline translation by introducing segment boundaries around every constituent phrase whose shallow syntactic parse category (from step 2) was V, NP, or PP. This breaks the baseline translation into a small number of coarse chunks to consider during re-ordering, instead of a large number of individual words.
5. Generate a set of candidate re-ordered translation hypotheses by iteratively moving constituent phrases whose predicate or semantic role label was *mismatched* to the input sentence. Each new candidate generated may in turn spawn a further set of candidates (especially since moving one constituent phrase may cause another's predicate or semantic role label to change from matched to mismatched). This search is performed breadth-first to favor fewer re-orderings (in case the hypothesis generation grows beyond allotted time).
6. Apply a semantic parser for the *output* language to each candidate re-ordered translation hypothesis as it is generated.
7. Return the re-ordered translation hypothesis with the maximum match of semantic predicates and arguments.

Figure 1. Algorithm for second pass.

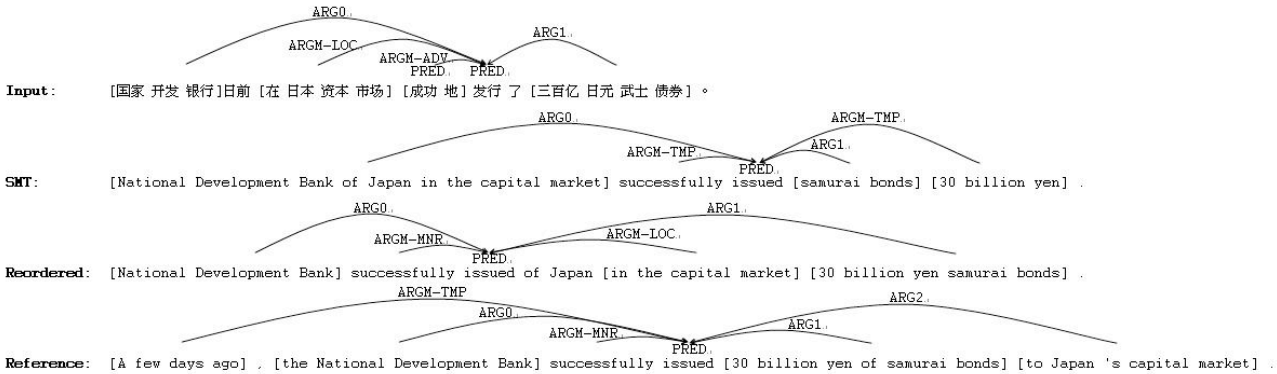

Figure 2. Example, showing translations after SMT first pass and after re-ordering second pass.

The first pass is performed using a conventional phrase-based SMT model. The phrase-based SMT model is assigned to the tasks of (a) providing an initial baseline hypothesis translation, and (b) fixing the lexical choice decisions. Note that the lexical choice decisions are *not* only at the single-word level, but are in general at the *phrasal* level.

The second pass takes the output of the first pass, and re-orders constituent phrases corresponding to semantic predicates and arguments, seeking to maximize the cross-lingual match of the semantic parse of the re-ordered translation to that of the original input sentence. The second pass algorithm performs the error correction shown in Figure 1.

The design decision to allow the first pass to fix all lexical choices follows an insight inspired by an empirical observation from our error analyses: the lexical choice decisions being made by today's SMT models have attained fairly reasonable levels, and are not where the major problems of adequacy lie. Rather, the ordering of arguments in relation to their predicates is often where the main failures of adequacy occur. By avoiding lexical choice variations while considering re-ordering hypotheses, a significantly larger amount of re-ordering can be done without further increasing computational complexity. So we sacrifice a small amount of fluency by allowing re-ordering without compensating lexical choice – in exchange for gaining potentially a larger amount of fluency by getting the predicate-argument structure right.

The model has a similar rationale for employing a re-ordering pass instead of re-ranking *n*-best lists or lattices. Oracle analysis of *n*-best lists and lattices show that they often focus on lexical choice alternatives rather than re-ordering / role variations which are more important to semantic adequacy.

3 Experiment

A Chinese-English experiment was conducted on the two-pass hybrid model. A phrase-based SMT baseline model was built by augmenting the open source statistical machine translation decoder Moses (Koehn *et al.* 2007) with additional pre-processors. English and Chinese shallow semantic parsers followed those discussed in Section 1.

The model was trained on LDC newswire parallel text consisting of 3.42 million sentence pairs, containing 64.1 million English words and 56.9 million Chinese words. The English was tokenized and case-normalized; the Chinese was tokenized via a maximum-entropy model (Fung *et al.* 2004).

Phrase translations were extracted via the grow-diag-final heuristic.

The language model is a 6-gram model trained with Kneser-Ney smoothing using the SRI language modeling toolkit (Stolcke 2002).

The test set of Wall Street Journal newswire sentences was randomly extracted from the Chinese-English Bilingual Propbank. Although we did not make use of the Propbank annotations, this would potentially allow other types of analyses in the future.

The phrase-based SMT model used for the first pass achieves a BLEU score of 42.99, establishing a fairly strong baseline to begin with.

In comparison, the automatically error-corrected translations that are output by the second pass achieve a BLEU score of 43.51. This represents approximately half a point improvement over the strong baseline.

An example is seen in Figure 2. The SMT first pass translation has an ARG0 *National Development Bank of Japan in the capital market* which is badly mismatched to *both* the input sentence's

ARG0 国家 开发 银行 and ARGM-LOC 在 日本 资本 市场. The second pass ends up re-ordering the constituent phrase corresponding to the mismatched ARGM-LOC, *of Japan in the capital market*, to follow the PRED *issued*, where the new English semantic parse now assigns most of its words the correctly matched ARGM-LOC semantic role label. Similarly, *samurai bonds 30 billion yen* is re-ordered to *30 billion yen samurai bonds*.

4 Discussion and conclusion

To our knowledge, this is a first result demonstrating that shallow semantic parsing can improve translation accuracy of SMT models. We note that accuracy here was measured via BLEU, and it has been widely observed that the negative impacts of semantic predicate-argument errors on the utility of the translation are underestimated by evaluation metrics based on lexical criteria such as BLEU. We conjecture that more expensive manual evaluation techniques which directly measure translation utility could even more strongly reveal improvement in role confusion errors.

The hybrid two-pass approach can be compared with the greedy re-ordering based strategy of the ReWrite decoder (Germann *et al.* 2001), although our search is breadth-first rather than purely greedy. Whereas ReWrite was based on word-level re-ordering, however, our approach is based on constituent phrase re-ordering, and the phrases to be re-ordered are more selectively chosen via the semantic parse labels. Moreover, the objective function being maximized by ReWrite is still the SMT model score; whereas in our case the new objective function is cross-lingual semantic predicate-argument match (plus an implicit search bias toward fewer re-orderings).

The hybrid two-pass approach can also be compared with serial combination architectures for hybrid MT (e.g., Ueffing *et al.* 2008). But whereas Ueffing *et al.* take the output from a first-pass rule-based MT system, and then correct it using a second-pass SMT system, our two-pass semantic SMT model does the reverse: it takes the output from a first-pass SMT system, and then corrects it with the aid of semantic analyzers.

Acknowledgments. Thanks to Chi-kiu Lo and Zhaojun Wu. 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.

References

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

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

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

Pascale Fung, Grace Ngai, Yongsheng Yang and Benfeng Chen. 2004. A maximum-entropy Chinese parser augmented by transformation-based learning. *ACM Transactions on Asian Language Information Processing (TALIP)* **3**(2): 159-168.

Pascale Fung, Zhaojun Wu, Yongsheng Yang and Dekai Wu. 2006. Automatic learning of Chinese/English semantic structure mapping. *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*. Aruba: Dec 2006. 230-233.

Pascale Fung, Zhaojun Wu, Yongsheng Yang and Dekai Wu. 2007. Learning Bilingual Semantic Frames: Shallow Semantic Parsing vs. Semantic Role Projection. *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*. Skövde, Sweden: Sep 2007. 75-84.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu and Kenji Yamada. Fast Decoding and Optimal Decoding for Machine Translation. 2001. *39th Annual Meeting of the Association for Computational Linguistics* (ACL-2001). Toulouse: July 2001.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic Labeling of Semantic Roles. *38th Annual Conference of the Association for Computational Linguistics (ACL-2000)*. 512–520, Hong Kong: Oct 2000.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. *WMT 2007 (ACL'07)*.

Jesús Giménez and Lluís Màrquez. 2007. Discriminative Phrase Selection for Statistical Machine Translation. *Learning Machine Translation*. NIPS Workshop Series. MIT Press.

Jesús Giménez and Lluís Màrquez. 2008. A Smorgasbord of Features for Automatic MT Evaluation. *3rd ACL Workshop on Statistical Machine Translation (shared evaluation task)*. Pages 195-198, Columbus, Ohio: Jun 2008.

Alessandro Moschitti and Roberto Basili. 2005. Verb subcategorization kernels for automatic semantic labeling. *ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Ann Arbor: Jun 2005. 10-17.

Franz Josef 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. 2004. A smorgasbord of features for statistical machine translation. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*. Boston: May 2004.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics* (ACL-2002).

Sameer Pradhan. 2005. *ASSERT: Automatic Statistical SEmantic Role Tagger*. http://oak.colorado.edu/assert/.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin and Daniel Jurafsky. 2005. Support Vector Learning for Semantic Argument Classification. *Machine Learning* **60**(1-3): 11-39.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin and Daniel Jurafsky. 2004. Shallow Semantic Parsing using Support Vector Machines. *Human Language Technology/North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*. Boston: May 2004.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. International Conference on Spoken Language Processing (ICSLP-2002). Denver, Colorado: Sep 2002.

Honglin Sun and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*. 249-256. Boston: May 2004.

Nicola Ueffing, Jens Stephan, Evgeny Matusov, Loïc Dugast, George Foster, Roland Kuhn, Jean Senellart and Jin Yang. 2008. Tighter Integration of Rule-based and Statistical MT in Serial System Combination. *22nd International Conference on Computational Linguistics (COLING 2008)*. Manchester: Aug 2008.

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

Dekai Wu and David Chiang (eds). 2009. *Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation* (NAACL-HLT 2009). Boulder, CO: Jun 2009.

Nianwen Xue and Martha Palmer. 2005. Automatic Semantic Role Labeling for Chinese Verbs. *19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland.