Semantic Cohesion Model for Phrase-based SMT

Minwei FENG[1] *Weiwei SUN*[2] *Hermann NEY*[1]
(1) Human Language Technology and Pattern Recognition Group,
Computer Science Department,
RWTH Aachen University,
Aachen, Germany
(2) The MOE Key Laboratory of Computational Linguistics,
Institute of Computer Science and Technology,
Peking University,
Beijing, China
feng@cs.rwth-aachen.de, ws@pku.edu.cn, ney@cs.rwth-aachen.de

ABSTRACT
In this paper, we propose a novel semantic cohesion model. Our model utilizes the predicate-argument structures as soft constraints and plays the role as a reordering model in the phrase-based statistical machine translation system. We build a translation system with GALE data. Experimental results on the NIST02, NIST03, NIST04, NIST05 and NIST08 Chinese-English tasks show that our model improves the baseline system by 0.93 BLEU 0.98 TER on average. We also compare our method with a syntax-augmented model (Cherry, 2008), and demonstrate the importance of predicate-argument semantics in machine translation.

KEYWORDS: statistical machine translation, semantic role labeling.

1 Introduction

In recent years, there are growing interests in incorporating semantics into statistical machine translation (SMT) (Wu and Fung, 2009; Liu and Gildea, 2010; Gao and Vogel, 2011; Baker et al., 2012). Among all existing semantic representations, semantic role labeling (SRL) aims at automatically analyzing predicate-argument structures and can capture essential meaning of sentences. Since the seminal work of (Gildea and Jurafsky, 2002), quite a few researchers concentrate on resolving SRL with different machine learning methods. Nowadays, good SRL systems can be built based on accurate syntactic parsers for English, as well as many other languages.

In this paper, we explore predicate-argument analysis of source sentences to improve phrase-based SMT systems. On one hand, the predicate-argument event layer of SRL captures global dependencies which is crucial for the MT output quality. On the other hand, the semantic role information contained in SRL also provide a good clue to the appropriateness of a phrase segment chosen by a translation system. Compared to the syntactic representation in both constituency and dependency formalisms, SRL focuses more on modeling the skeleton of a sentence.

To exploit the two attributes of SRL, we propose two types of constraints which are implemented as two models in the decoder. The first model restrains the translation process so that it is consistent with the global dependency in the source sentence. The second model inspects the source sentence segmentation so that each source phrase is consistent with the semantic roles.

We conduct experiments on the NIST02, NIST03, NIST04, NIST05 and NIST08 Chinese-English tasks. Experimental results show that our model improves the baseline system by 0.93 BLEU 0.98 TER on average. We also compare our method with a syntax-augmented model (Cherry, 2008), and demonstrate the importance of predicate-argument semantics in machine translation.

The remainder of this paper is organized as follows: Section 2 describes the semantic soft constraints proposed for translation system. Section 3 introduces a related work (Cherry, 2008) as we compare our method with it. Section 4 provides the experimental configuration and results. Section 5 briefly summarizes recent work on employing different semantics related techniques. Conclusions will be given in Section 6.

2 Semantic Cohesion Model

In this section, we describe the proposed idea. Firstly, we will briefly describe the basement of this research: the principle of statistical machine translation. Secondly, the SRL framework will be given. Thirdly, we demonstrate how the semantic role information can be used for translation.

2.1 Principle

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$. The objective is to translate the source into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. The strategy is among all possible target language sentences, we will choose the one with the highest probability:

$$\hat{e}_i^{\hat{I}} = \arg\max_{I, e_1^I} \{Pr(e_1^I | f_1^J)\} \quad (1)$$

We model $Pr(e_1^I|f_1^J)$ directly using a log-linear combination of several models (Och and Ney, 2002):

$$Pr(e_1^I|f_1^J) = \frac{\exp\Big(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\Big)}{\sum_{I', e'^{I'}_1} \exp\Big(\sum_{m=1}^{M} \lambda_m h_m(e'^{I'}_1, f_1^J)\Big)} \tag{2}$$

The denominator is to make the $Pr(e_1^I|f_1^J)$ to be a probability distribution and it depends only on the source sentence f_1^J. For search, the decision rule is simply:

$$\hat{e}_i^{\tilde{I}} = \arg\max\Big\{\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\Big\} \tag{3}$$

The model scaling factors λ_1^M are trained with Minimum Error Rate Training (MERT).

In this paper, the phrase-based machine translation system is utilized (Och et al., 1999; Zens et al., 2002; Koehn et al., 2003). The translation process consists in segmentation of the source sentence according to the phrase table which is built from the word alignment. The translation of each of these segments is then just extracting the target side from the phrase pair. With the corresponding target side, the final translation is the composition of these translated segments. In this last step, reordering is allowed.

2.2 Semantic Role Labeling

In the last decade, there has been an increasing interest in SRL on several languages, which consists of recognizing arguments involved by predicates in a given sentence and labeling their semantic types. Typical semantic classes include *Agent*, *Patient*, *Source*, *Goal*, and so forth, which are core arguments to a predicate, as well as *Location*, *Time*, *Manner*, *Cause*, and so on, which are adjuncts. In order to indicate exactly what semantic relations hold among a given predicate and its associated participants and properties, the role-bearing constituents must be identified and their correct semantic role labels assigned.

Such sentence-level semantic analysis of text is concerned with the characterization of events and is therefore important to understand the essential meaning of the original input language sentences – *who* did *what* to *whom*, for *whom* or *what*, *how*, *where*, *when* and *why*? Different from many other semantic representation formalisms, this shallow semantic interpretation abstracts important predicate-argument structural information away from syntactic structure and may potentially benefit machine translation, as well as many other NLP applications.

2.3 Semantic Cohesion for Translation

We now explain how the SRL is utilized in the decoder. Suppose the source language is temporarily English, we detect and classify semantic roles of all predicates for the source sentence before translation. The Figure 1 is an example with Propbank-style annotations. Figure 1 consists of four rows. The first row is the orginal source sentence. The next three rows demonstrate the event layers of the sentence. Each event layer has one predicate and corresponding arguments. The source sentence in Figure 1 has three predicates. The first predicate *make* and second predicate *distribute* governs two arguments A0 (namely proto-Agent) and A1 (namely proto-Patient) respectively; third predicate *base* possesses two semantic

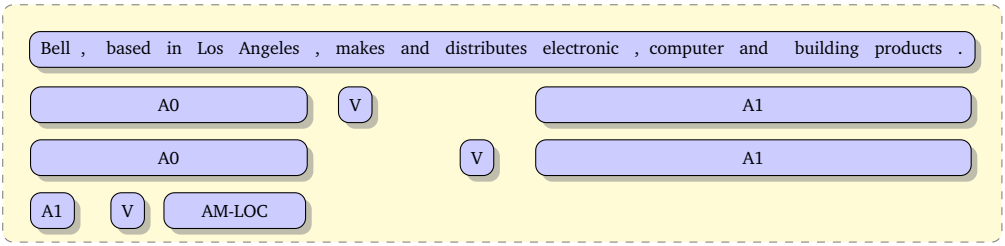


Figure 1: An example of SRL. The source sentence is given at top.

roles A1 and AM-LOC (namely location). Then a model will be added into the decoder as a new feature of the log-linear framework (Equation 2). The algorithm is as follows:

Given the source sentence and its predicate-argument structural information, during the translation process, every time one hypothesis is extended, the added model checks if the source semantic analysis contains one structure S such that:

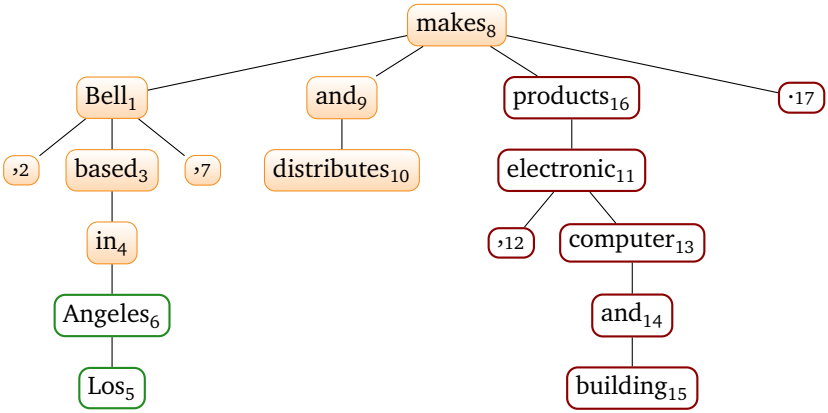
- Its translation is already started (at least one word is covered)
- It is interrupted by the new added phrase (at least one word in the new source phrase is not in S)
- It is not finished (after the new phrase is added, there is still at least one uncovered source word in S)

If so, we say this hypothesis violates the structure S, and the model returns the number of structures that this hypothesis violates.

We use two kinds of structures. In other words, we add two models/features into the log-linear framework. The structure of the first model (we name it SRL1) is the whole predicate-argument structure, i.e. event layer. The SRL1 feature will report how many event layers that one search state violates. For the second model (we name it SRL2), the structure is semantic role. The SRL2 feature will calculate the amount of semantic roles that one search state violates. In Figure 1, there are three event layers and six semantic roles. Suppose currently only the first source word *Bell* is already covered/translated, and then the decoder decides to translate the source word *computer*. Then the feature SRL1 will give penalty 1. Because this decision violates the third event layer in Figure 1. The third event layer was starting to be translated (*Bell* is covered), however, the decoder jumps to *computer* before it finishes current event layer(A1 and AM-LOC have not been translated yet). The feature SRL2 will give penalty 2 because the two semantic role A0 have been violated. During search, we have the access to the coverage vector which stores those source words that have been translated. So based on the source sentence semantic analysis and current coverage vector, the two feature values can be easily computed.

3 Syntactic cohesion model

We now introduce a related work as comparison will be presented later in this paper. (Cherry, 2008) proposed a syntactic cohesion model. The core idea is that the syntactic structure of the source sentence should be preserved during translation. This structure is represented by a source sentence dependency tree. To keep syntactic cohesion, the decoding process should not break



[Bell , based in Los Angeles , makes and distributes electronic , computer and building products .]

Figure 2: A dependency tree example. The source sentence is given at bottom.

this dependency structure. (Cherry, 2008) used his model as a new feature of the log-linear decoding framework and showed improvement on English-French direction. We implement this model in the phrase-based decoder and report results on Chinese-English translation.

To illustrate the method, we use the Figure 2 as an example. Figure 2 is a dependency tree (again, suppose the source language is now English) . Every node represents a word. We also put the position of the word in the node. So $Bell_1$ means $Bell$ is the first word of the sentence. The algorithm is as follows:

Given the source sentence and its dependency tree, during the translation process, once a hypothesis is extended, check if the source dependency tree contains a subtree T such that:

- Its translation is already started (at least one node is covered)
- It is interrupted by the new added phrase (at least one word in the new source phrase is not in T)
- It is not finished (after the new phrase is added, there is still at least one free node in T)

If so, we say this hypothesis violates the subtree T, and the model returns the number of subtrees that this hypothesis violates.

In Figure 2, nodes filled with yellow means the words have been already covered/translated. Now suppose the length of the new added phrase is 1, then according to the above algorithm only position 5 and 6 (green rectangle) are good candidates. Choosing other source words (red rectangle) to translate will violate the subtree $in_4 - Los_5 - Angeles_6$.

4 Experiments

4.1 Experimental Setup

Our baseline is a phrase-based decoder, which includes the following models: an n-gram target-side language model, a phrase translation model and a word-based lexicon model. The

latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally we use phrase count features, word and phrase penalty. The reordering model for the baseline system is the distance-based jump model which uses linear distance. This model does not have hard limit. We list the important information regarding the experimental setup below. All those conditions have been kept same in this work.

- lowercased training data (Table 1) from GALE task
  alignment trained with GIZA++
- tuning corpus: NIST06 test corpora: NIST02 03 04 05 and 08
- 5-gram LM (1 694 412 027 running words) trained by SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing
  LM training data: target side of bilingual data.
- BLEU (Papineni et al., 2001) and TER (Snover et al., 2005) reported
  all scores calculated in lowercase way.
- Stanford Parser (Levy and Manning, 2003) used to get the Chinese constituent tree for the SRL and the dependency tree for the syntactic cohesion model

	Chinese	**English**
Sentences	5 384 856	
Running Words	115 172 748	129 820 318
Vocabulary	1 125 437	739 251

Table 1: training data statistics

4.2 A Full Parsing Based Chinese SRL System

4.2.1 Background

SRL methods that are successful on English are adopted to resolve Chinese SRL (Xue, 2008; Sun, 2010). Previous work indicates that syntactic information is very important for SRL and full parsing based approaches are considerably better than shallow parsing based ones. Based on a phrase-structure parsing, SRL is usually formulated as a constituent classification problem. In particular, SRL is divided into three sub-tasks: 1) pruning with a heuristic rule, 2) argument identification (AI) to recognize arguments, and 3) semantic role classification (SRC) to predict semantic types. To efficiently excluded non-arguments, a pruning procedure is executed to filter out constituents that are highly unlikely to be semantic roles of a predicate p. In the AI sub-task, for every candidate constituent c, a binary classifier is employed to determine whether c is an argument of p. In the SRC sub-task, all arguments recognized in the AI step are assigned detailed semantic types by a multi-class classifier. Distinct from the constituent classification method introduced above, we use a constituent chunking based method to acquire predicate-argument structures in this paper.

4.2.2 Semantic Chunking Based SRL

SRL is performed on the output of a syntactic parser, and only phrases in the parse tree are taken as possible candidates. If there is no phrase in the parse tree that shares the same text

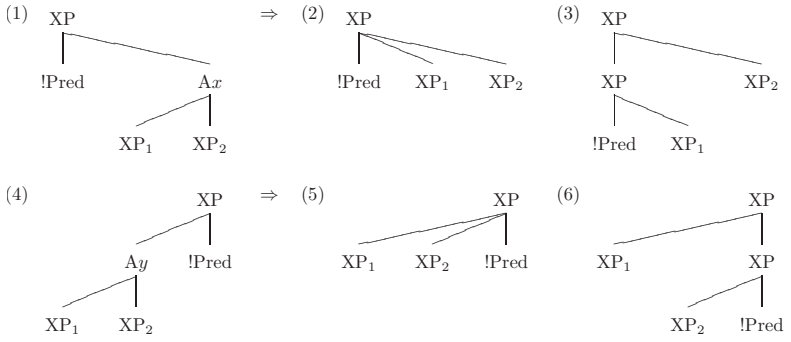


Figure 3: Parsing errors that can be tolerated by full parsing based constituent chunking.

span with an argument in the manual annotation, the system cannot possibly get a correct prediction. In other words, the best the system can do is to correctly label all arguments that have a counterpart node in the parse tree.

In this paper, we implement a semantic chunking method which can tolerate some syntactic parsing errors. First, our system collects all c-commanders[1] and puts them in order. Because c-commanders of a predicate are not overlapped with each other and compose the whole sentence, we can take this step as a sequentialization procedure. Sun et al. (2008) present a theoretical analysis about argument positions and suggest that an argument should c-command a predicate. Therefore, our sequentialization precedure keeps most semantic roles. On basis of sequentialized constituents, we define semantic chunks which do not overlap nor embed using IOB2 representation (Ramshaw and Marcus, 1995) and transfer the SRL problem as a **constituent tagging** problem. Our definition of semantic chunks is described below.

- Constituent outside an argument receive the tag *O*.
- For a sequence of constituents forming a semantic role of *Ax*, the first constituent receives the semantic chunk label *B(egin)-Ax*,
- and the remaining ones receive the label *I(nside)-Ax*.

Developing features has been shown crucial to advancing the state-of-the-art in SRL. To achieve good Chinese SRL results, we utilize rich syntactic features introduced in (Sun, 2010). For sequential tagging, we use a first order linear-chain global linear model and estimate parameters with structured preceptron (Collins, 2002).

Our semantic chunking method can tolerate two types of parsing errors that are shown in Figure 3. Assume tree structures (1 and 4) on the left hand side are the correct syntactic analysis, while tree structures (2, 3, 5 and 6) on the right hand side are some wrong analysis. Though a constituent classification system, the arguments Ax and Ay can not be recovered since there is no node to express them. In our constituent chunking system, however, when these errors occur, the arguments can still be found, if XP_1 is assigned a label *B-Ax* or *B-Ay* and XP_2 is assigned a label *I-Ax* or *I-Ay*.

[1]C-command is an concept in X-bar theory. Assuming α and β are two nodes in a syntax tree: α C-commands β means every parent of α is ancestor of β.

The full parsing based semantic chunking system was first introduced in (Sun, 2012). To learn more empirical evaluation results of this system as well as a comparative study of full and shallow parsing based SRL, readers can refer to (Sun, 2012).

4.2.3 An Example

An example is illustrated in Figure 4, where the predicate is the verb "调查/investigate". To find all arguments and adjuncts, our system first employs Stanford parser to obtain a phrase-structure analysis. Based on the syntactic tree, our SRL system then collects all c-commanders of the predicate, including all square nodes. Finally, a sequence classifier is applied to assign these candidate nodes chunk labels *B-A0*, *B-TMP*, *B-MNR* and *B-A1*. Based on these labels, we can know the event structure, e.g. the proto-Patient is the NP "事故原因/accident cause".

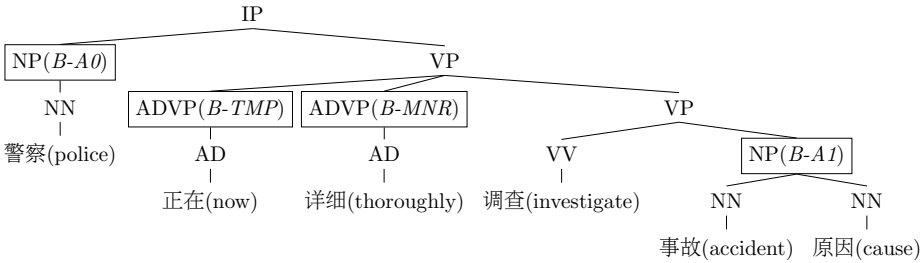


Figure 4: An example sentence: *The police are thoroughly investigating the cause of the accident*.

4.3 Results

Experimental results are presented in Table 2. Besides the five test corpora, we add a column **avg.** to show the average improvements. We also add a column **Index** for score reference convenience. **SRL1** and **SRL2** are the two features described in Section 2.2. **SC** is the syntactic cohesion model described in Section 3.

From Table 2 we see that our proposed feature **SRL1** is able to improve the baseline by 0.57 BLEU and 0.71 TER. **SRL2** improves the baseline by 0.75 BLEU and 0.77 TER. When **SRL1** and **SRL2** are used together, further improvements have been observed. In general, the semantic analysis information improves the baseline system by 0.93 BLEU and 0.98 TER (line 4 and 9). For the comparison with **SC** model, we see their performance is very close. Both **SRL1+SRL2** and **SC** refine the baseline. **SRL1+SRL2** is slightly better at BLEU (compare line 4 and 5) while **SC** is slightly better with TER (compare line 9 and 10).

SRL abstracts important event structures away from syntactic parses. Compared to full parsing model **SC**, **SRL1+SRL2** only concentrates on modeling the skeleton of a sentence, and provide much less information. Nevertheless, our experiments suggest that SRL achieves an equivalent contribution (as constraints) to a phrase-based MT system.

5 Previous Work

(Wu and Fung, 2009) propose a Hybrid two-pass model to use semantics for SMT. The first pass is a conventional phrase-based SMT decoder. The second pass generate a set of candidate re-ordered translation hypotheses by iteratively moving constituent phrases whose predicate or semantic role label was mismatched to the source sentence. The output of the second pass

Systems	NIST02	NIST03	NIST04	NIST05	NIST08	avg.	Index
			BLEU scores				
baseline	33.60	34.29	35.73	32.15	26.34	-	1
baseline+SRL1	34.50	34.76	36.21	32.75	26.77	0.57	2
baseline+SRL2	34.73	34.88	36.60	32.83	26.82	0.75	3
baseline+SRL1+SRL2	35.05	34.93	36.71	33.22	26.89	0.93	4
baseline+SC	34.96	34.52	36.37	33.35	26.90	0.79	5
			TER scores				
baseline	61.36	60.48	59.12	60.94	65.17	-	6
baseline+SRL1	60.44	59.58	58.61	60.01	64.88	0.71	7
baseline+SRL2	60.28	59.49	58.14	60.20	65.11	0.77	8
baseline+SRL1+SRL2	60.05	59.55	58.14	59.69	64.74	0.98	9
baseline+SC	59.90	59.37	58.27	59.69	64.44	1.08	10

Table 2: Experimental results

is the re-ordered translation hypothesis with the maximum match of semantic predicates and arguments. In essence, the paper proposes a semantics-based postprocessing or a semantics-based post reordering technique, where the usage of semantic information is limited due to the fact that the lexical choice has been fixed. The second pass can only do some permutation of the output of the phrase-based decoder.

(Liu and Gildea, 2010) implement two semantic role features in their tree-to-string machine translation system. First feature is named semantic role reordering which describes reordering of the source side semantic roles in the target side. Second feature is named deleted role which can penalize the deletion of source side semantic roles. They show improvement over a small FBIS English-to-Chinese system.

(Gao and Vogel, 2011) present an approach of utilizing target side SRL information to improve the hierarchical phrase-based machine translation. They extract SRL-aware Synchronous Context-Free Grammar (SCFG) rules together with conventional Hiero rules. Instead of adding additional features in the decoder, special conversion rules are applied during rule extraction procedure to ensure that when SRL-aware SCFG rules are used in derivation, the decoder only generates hypotheses with complete semantic structures.

(Baker et al., 2012) build a modality/negation (MN) annotation scheme in the target side. They define modality as an extra-propositional component of meaning and negation as an inextricably intertwined component of modality. A MN lexicon is created in a semi-supervised way. Using this lexicon a MN tagger is designed to identify the substrings that are related to MN. After the target side MN annotation, they incorporate the MN into a small Urdu-English translation task. Their baseline is a Syntax Augmented Machine Translation (SAMT) system. By using a tree grafting approach, the original syntactic tags enriched with new MN annotations are assigned to the parse trees of target side training data.

The main characteristics (also the differences to the above work) of this paper are as follows: we utilize the source side SRL information as soft constraints to improve the phrase-based machine translation system; the two models/features are implemented in the log-linear framework so that SRL information can directly act on the translation process; the experiments are carried out with a large scale Chinese-to-English NIST task; a comparison with (Cherry, 2008) which

uses dependency tree as soft constraints has been conducted.

6 Conclusion

In this paper, we proposed a method to utilize the source side semantic analysis for SMT. Two models have been created. The first model SRL1 captures the global semantic dependency in the source sentence; the second model SRL2 makes sure that the local semantic coherence is kept during search. From SMT perspective, SRL1 is a long distance reordering model; SRL2 is a local reordering model and a phrase model (it encourages the decoder to choose those phrases that keep semantic coherence). The SRL is able to improve the baseline 0.93 BLEU and 0.98 TER. We also did comparative study with SC model. Results show that the performance of the two methods is similar. SRL is 0.14 BLEU better than SC while SC is 0.1 TER better than SRL.

Acknowledgments

This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 4911028154.0. 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 (DARPA). This work was also partly supported by National High-Tech R&D Program (2012AA011101).

The authors would like to give special thanks to anonymous reviewers for their valuable comments and suggestions.

References

Baker, K., Bloodgood, M., Dorr, B. J., Callison-Burch, C., Filardo, N. W., Piatko, C., Levin, L., and Miller, S. (2012). Modality and negation in simt use of modality and negation in semantically-informed syntactic mt. *Comput. Linguist.*, 38(2):411–438.

Cherry, C. (2008). Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 72–80, Columbus, Ohio.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics.

Gao, Q. and Vogel, S. (2011). Utilizing target-side semantic role labels to assist hierarchical phrase-based machine translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, pages 107–115, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Levy, R. and Manning, C. D. (2003). Is it harder to parse chinese, or the chinese treebank? In *Proceedings of ACL-03*, pages 439–446, Sapporo, Japan.

Liu, D. and Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 716–724, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL-02*, pages 295–302, Philadelphia, Pennsylvania, USA.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, pages 20–28, University of Maryland, College Park, MD, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. (RC22176 (W0109-022)).

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In Yarowsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., and Weischedel, R. (2005). A study of translation error rate with targeted human annotation. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, MD.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings of ICSLP-02*, pages 901–904, Denver, Colorado,USA.

Sun, W. (2010). Improving Chinese semantic role labeling with rich syntactic features. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 168–172, Uppsala, Sweden. Association for Computational Linguistics.

Sun, W. (2012). *Learning Chinese Language Structures with Multiple Views*. PhD thesis, Saarland University.

Sun, W., Sui, Z., and Wang, H. (2008). Prediction of maximal projection for semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 833–840, Manchester, UK. Coling 2008 Organizing Committee.

Wu, D. and Fung, P. (2009). Semantic roles for smt: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xue, N. (2008). Labeling Chinese predicates with semantic roles. *Computational Linguistics*, 34:225–255.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *German Conference on Artificial Intelligence*, pages 18–32. Springer Verlag.