Comparing Representations of Semantic Roles for String-To-Tree Decoding

Marzieh Bazrafshan and **Daniel Gildea**
Department of Computer Science
University of Rochester
Rochester, NY 14627

Abstract

We introduce new features for incorporating semantic predicate-argument structures in machine translation (MT). The methods focus on the completeness of the semantic structures of the translations, as well as the order of the translated semantic roles. We experiment with translation rules which contain the core arguments for the predicates in the source side of a MT system, and observe that using these rules significantly improves the translation quality. We also present a new semantic feature that resembles a language model. Our results show that the language model feature can also significantly improve MT results.

1 Introduction

In recent years, there have been increasing efforts to incorporate semantics in statistical machine translation (SMT), and the use of predicate-argument structures has provided promising improvements in translation quality. Wu and Fung (2009) showed that shallow semantic parsing can improve the translation quality in a machine translation system. They introduced a two step model, in which they used a semantic parser to rerank the translation hypotheses of a phrase-based system. Liu and Gildea (2010) used semantic features for a tree-to-string syntax based SMT system. Their features modeled deletion and reordering for source side semantic roles, and they improved the translation quality. Xiong et al. (2012) incorporated the semantic structures into phrase-based SMT by adding syntactic and semantic features to their translation model. They proposed two discriminative models which included features for predicate translation and argument reordering from source to target side. Bazrafshan and Gildea (2013) used semantic structures in a string-to-tree translation system by extracting translation rules enriched with semantic information, and showed that this can improve the translation quality. Li et al. (2013) used predicate-argument structure reordering models for hierarchical phrase-based translation, and they used linguistically motivated constraints for phrase translation.

In this paper, we experiment with methods for incorporating semantics in a string-to-tree MT system. These methods are designed to model the order of translation, as well as the completeness of the semantic structures. We extract translation rules that include the complete semantic structure in the source side, and compare that with using semantic rules for the target side predicates. We present a method for modeling the order of semantic role sequences that appear spread across multiple syntax-based translation rules, in order to overcome the problem that a rule representing the entire semantic structure of a predicate is often too large and too specific to apply to new sentences during decoding. For this method, we compare the verb-specific roles of PropBank and the more general thematic roles of VerbNet.

These essential arguments of a verbal predicate are called the *core* arguments. Standard syntax-based MT is incapable of ensuring that the target translation includes all of the core arguments of a predicate that appear in the source sentence. To encourage the translation of the likely core arguments, we follow the work of Bazrafshan and Gildea (2013), who use special translation rules with complete semantic structures of the predicates in the target side of their MT system. Each of these rules includes a predicate and all of its core arguments. Instead of incorporating only the target side semantic rules, we extract the special rules for both the source and the target sides, and compare the effectiveness of adding these rules to

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1786–1791,
October 25-29, 2014, Doha, Qatar.

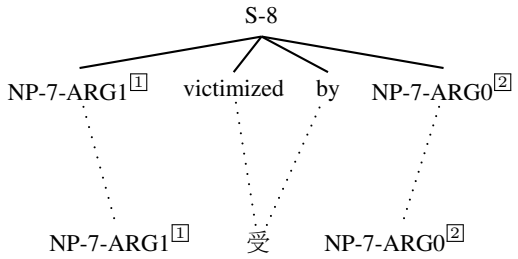

Figure 1: A complete semantic rule (Bazrafshan and Gildea (2013)).

the system separately and simultaneously.

Besides the completeness of the arguments, it is also important for the arguments to appear in the correct order. Our second method is designed to encourage correct order of translation for both the core and the non-core roles in the target sentence. We designed a new feature that resembles the language model feature in a standard MT system. We train a n-gram language model on sequences of semantic roles, by treating the semantic roles as the *words* in what we call the *semantic language*. Our experimental results show that the language model feature significantly improves translation quality.

Semantic Role Labeling (SRL): We use semantic role labelers to annotate the training data that we use to extract the translation rules. For target side SRL, the role labels are attached to the nonterminal nodes in the syntactic parse of each sentence. For source side SRL, the labels annotate the spans from the source sentence that they cover. We train our semantic role labeler using two different standards: Propbank (Palmer et al., 2005) and VerbNet (Kipper Schuler, 2005).

PropBank annotates the Penn Treebank with predicate-argument structures.It use generic labels (such as Arg0, Arg1, etc.) which are defined specifically for each verb. We trained a semantic role labeler on the annotated Penn Treebank data and used the classifier to tag our training data.

VerbNet is a verb lexicon that categorizes English verbs into hierarchical classes, and annotates them with thematic roles for the arguments that they accept. Since the thematic roles use more meaningful labels (e.g. Agent, Patient, etc.), a language model trained on VerbNet labels may be more likely to generalize across verbs than one trained on PropBank labels. It may also provide more information, since VerbNet has a larger set of labels than PropBank. To train the semantic role labeler on VerbNet, we used the mappings

$$\begin{array}{rl} A \rightarrow BC & c_0 \\ \quad [B,i,j] & c_1 \\ \quad [C,j,k] & c_2 \\ \hline \quad [A,i,k] & c_0 + c_1 + c_2 \end{array}$$

Figure 2: A deduction step in our baseline decoder

provided by the SemLink project (Palmer, 2009) to annotate the Penn Treebank with the VerbNet roles. These mappings map the roles in PropBank to the thematic roles of VerbNet. When there is no mapping for a role, we keep the role from Propbank.

2 Using Semantics in Machine Translation

In this section, we present our techniques for incorporating semantics in MT: source side semantic rules, and the semantic language model.

2.1 Source Side Semantic Rules

Bazrafshan and Gildea (2013) extracted translation rules that included a predicate and all of its arguments from the target side, and added those rules to the baseline rules of their string-to-tree MT system. Figure 1 shows an example of such rules, which we refer to as *complete semantic rules*. The new rules encourage the decoder to generate translations that include all of the semantic roles that appear in the source sentence.

In this paper, we use the same idea to extract rules from the semantic structures of the source side. The complete semantic rules consist of the smallest fragments of the combination of GHKM (Galley et al., 2004) rules that include one predicate and all of its core arguments that appear in the sentence. Rather than keeping the predicate and argument labels attached to the non-terminals, we remove those labels from our extracted semantic rules, to keep the non-terminals in the semantic rules consistent with the non-terminals of the baseline GHKM rules. This is also important when using both the source and the target semantic rules (i.e. Chinese and English rules), as it has been shown that there are cross lingual mismatches between Chinese and English semantic roles in bilingual sentences (Fung et al., 2007).

We extract a complete semantic rule for each verbal predicate of each sentence pair in the training data. To extract the target side complete semantic rules, using the target side SRL anno-

$A \rightarrow BC$ *to space*	c_0 (x1 x2 Destination)
$[B, i, j, (Agent,)]$	c_1
$[C, j, k, (PRED\_bring, Theme,)]$	c_2
$[A, i, k,$ (Agent, PRED_bring,-*-, Theme, Destination)$]$	$c_0 + c_1 + c_+$ + LMcost(Agent, PRED_bring,-*-, Theme, Destination)

Figure 3: A deduction step in the semantic language model method.

tated training data, we follow the general GHKM method, and modify it to ensure that each *frontier node* (Galley et al., 2004) in a rule includes either all or none of the semantic role labels (i.e. the predicate and all of its present core arguments) in its descendants in the target side tree. The resulting rule then includes the predicate and all of its arguments. We use the source side SRL annotated training data to extract the source side semantic rules. Since the annotations specify the spans of the semantic roles, we extract the semantic rules by ensuring that the span of the root (in the target side) of the extracted rule covers all of the spans of the roles in the predicate-argument structure.

The semantic rules are then used together with the original GHKM rules. We add a binary feature to distinguish the semantic rules from the rest. We experiment with adding the semantic rules from the source side, and compare that with adding semantic rules of both the source and the target side.

In all of the experiments in this paper, we use a string-to-tree decoder which uses a CYK style parser (Yamada and Knight, 2002). Figure 2 depicts a deduction step in the baseline decoder. The CFG rule in the first line is used to generate a new item A with span (i, k) using items B and C, which have spans (i, j) and (j, k) respectively. The cost of each item is shown on the right. For experimenting with complete semantic rules, in addition having more rules, the only other modification made to the baseline system is extending the feature vector to include the new feature. We do not modify the decoder in any significant way.

2.2 Semantic Language Model

The semantic language model is designed to encourage the correct order of translation for the semantic roles. While the complete translation rules of Section 2.1 contain the order of the translation for core semantic roles, they do not include the non-core semantic roles, that is, semantic roles which are not essential for the verbal predicates, but do contribute to the meaning of the predicate. In addition, the semantic LM can help in cases where no specific complete semantic rule can apply, which makes the system more flexible.

The semantic language model resembles a regular language model, but instead of words, it defines a probability distribution over sequences of semantic roles. For this method we also use a semantic role labeler on our training data, and use the labeled data to train a tri-gram semantic language model.

The rules are extracted using the baseline rule extraction method. As opposed to the previous method, the rules for this method are not derived by combining GHKM rules, but rather are regular GHKM rules which are annotated with semantic roles. We make a new field in each rule to keep the ordered list of the semantic roles in that rule. We also include the nonterminals of the right-hand-side of the rule in that ordered list, to be able to substitute the semantic roles from the input translation items in the correct order. The decoder uses this new field to save the semantic roles in the translation items, and propagates the semantic LM *states* in the same way that the regular language model states are propagated by the decoder.

We define a new feature for the semantic language model, and score the semantic states in each translation item, again analogously to a regular language model. Figure 3 depicts how the deduction for this method is different from our baseline. In this example, the semantic roles "Agent", "PRED_bring" and "Theme" come from the input items, and the role "Destination" (which tags the terminals "to space") comes from the translation rule.

We stemmed the verbs for training this feature, and also annotated our rules with stemmed verbal predicates. The stemming helps the training since the argument types of a verb are normally independent of its inflected variants.

	avg. BLEU Score		
	dev	test	p-value
Baseline	26.01	25.00	-
Source	26.44	25.17	0.048
Source and target	26.39	**25.63**	$< 10^{-10}$
Propbank LM	26.38	25.08	0.108
VerbNet LM	26.58	**25.23**	0.025

Table 1: Comparisons of the methods with the baseline. The BLEU scores are calculated on the top 3 results from 15 runs MERT for each experiments. The p-values are calculated by comparing each method against the baseline system.

3 Experiments

3.1 Experimental Setup

The data that we used for training the MT system was a Chinese-English corpus derived from newswire text from LDC.[1] The data consists of 250K sentences, which is 6.3M words in the English side. Our language model was trained on the English side of the entire data, which consisted of 1.65M sentences (39.3M words). Our development and test sets are from the newswire portion of NIST evaluations (2004, 2005, 2006). We used 392 sentences for the development set and 428 sentences for the test set. These sentences have lengths smaller than 30, and they each have 4 reference translations. We used our in-house string-to-tree decoder that uses Earley parsing. Other than the features that we presented for our new methods, we used a set of nine standard features. The rules for the baseline system were extracted using the GHKM method. Our baseline GHKM rules also include composed rules, where larger rules are constructed by combining two levels of the regular GHKM rules. We exclude any unary rules (Chung et al., 2011), and only keep rules that have scope up to 3 (Hopkins and Langmead, 2010). For the semantic language model, we used the SRILM package (Stolcke, 2002) and trained a tri-gram language model with the default Good-Turing smoothing.

Our target side semantic role labeler uses a maximum entropy classifier to label parsed sentences. We used Sections 02-22 of the Penn TreeBank to train the labeler, and sections 24 and 23 for development set and training set respectively. The labeler has a precision of 90% and a recall of 88%. We used the Chinese semantic role labeler of Wu and Palmer (2011) for source side SRL, which uses the LIBLINEAR (Fan et al., 2008) as a classifier. Minimum Error Rate Training (MERT) (Och, 2003) was used for tuning the feature weights. For all of our experiments, we ran 15 instances of MERT with random initial weight vectors, and used the weights of the top 3 results on the development set to test the systems on the test set. We chose to use the top 3 runs (rather than the best run) of each system to account for the instability of MERT (Clark et al., 2011). This method is designed to reflect the average performance of the MT system when trained with random restarts of MERT: we wish to discount runs in which the optimizer is stuck in a poor region of the weight space, but also to average across several good runs in order not to be mislead by the high variance of the single best run. For each of our MT systems, we merged the results of the top 3 runs on the test set into one file, and ran a statistical significance test, comparing it to the merged top 3 results from our baseline system. The 3 runs were merged by duplicating each run 3 times, and arranging them in the file so that the significance testing compares each run with all the runs of the baseline. We performed significance testing using paired bootstrap resampling (Koehn, 2004). The difference is considered statistically significant if $p < 0.05$ using 1000 iterations of paired bootstrap resampling.

3.2 Results

Our results are shown in Table 1. The second and the third columns contain the average BLEU score (Papineni et al., 2002) on the top three results on the development and test sets. The fourth column is the p-value for statistical significance testing against the baseline. The first row shows the results for our baseline. The second row contains the results for using the source (Chinese) side complete semantic rules of Section 2.1, and the third row is the results for combining both the source and the target side complete semantic rules. As noted before, in both of these experiments we also use the regular GHKM rules. The result show that the source side complete semantic rules improve the system ($p = 0.048$), and as we expected, combining the source and the tar-

[1]The data was randomly selected from the following sources: LDC2006E86, LDC2006E93, LDC2002E18, LDC2002L27, LDC2003E07, LDC2003E14, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E26, LDC2005E83, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006E24, LDC2006E92, LDC2006E24

Source Sentence	因此 , 保护 儿童 免 受 武装冲突 的 伤害 是 国际 社会 重要 的 职责 .
Reference	therefore , it is the international community 's responsibility to protect the children from harms resulted from armed conflicts .
Baseline	the armed conflicts will harm the importance of the international community the responsibilities . therefore , from child protection
Verbet LM	therefore , the importance of the international community is to protect children from the harm affected by the armed conflicts .

Source Sentence	同 去年 的 会议 相比 , 今年 会议 的 火药 味 消失 了 , 双方 的 立场 在 靠近 .
Reference	compared with last year 's meeting , the smell of gunpowder has disappeared in this year 's meeting and the two sides ' standpoints are getting closer .
Baseline	disappears on gunpowder , near the stance of the two sides compared with last year 's meeting , the meeting of this year .
Verbet LM	the smells of gunpowder has disappeared , the position in the two sides approach . compared with last year 's meeting , this meeting

(a) Comparison of the language model method (using VerbNet) and the baseline system.

Source Sentence	科学家 曾 大胆 预料 , 这 艘 英国 的 太空 船 可能 陷 在 坑 洞 中 .
Reference	scientists have boldly predicted that the british spacecraft might have been stuck in a hole .
Baseline	scientists boldly expected , this vessel uk may have in the space ship in hang tung .
Semantic Rules	scientists have boldly expected this vessel and the possible settlement of the space ship in hang tung .

Source Sentence	美国 政府 应 以 善意 对待 朝鲜 的 这 一 立场 .
Reference	the us government should show goodwills to north korea 's stand .
Baseline	this position of the government of the united states to goodwill toward the dprk .
Semantic Rules	this position that the us government should use goodwill toward the dprk .

(b) Comparison of the experiments with source and target side semantic rules and the baseline system.

Figure 4: Comparison of example translations from our semantic methods and the baseline system.

get side rules improves the system even more significantly ($p < 10^{-10}$). To measure the effect of combining the rules, in a separate experiment we replicated the complete semantic rules experiments of Bazrafshan and Gildea (2013), and ran statistical significance tests comparing the combination of the source and target rules with using only the source or the target semantic rules separately. The results showed that combining the semantic rules outperforms both of the experiments that used rules from only one side (with $p < 0.05$ in both cases).

The results for the language model feature are shown in the last two rows of the table. Using Propbank for language model training did not change the system in any significant way ($p = 0.108$), but using VerbNet significantly improved the results ($p = 0.025$). Figure 4(a) contains an example comparing the baseline system with the VerbNet language model. We can see how the VerbNet language model helps the decoder translate the argument in the correct order. The baseline system has also generated the correct arguments, but the output is in the wrong order. Figure 4(b) compares the experiment with semantic rules of both target and source side and the baseline system. Translation of the word "use" by our semantic rules is a good example showing how the decoder uses these semantic rules to generate a more complete predicate-argument structure.

4 Conclusions

We experimented with two techniques for incorporating semantics in machine translation. The models were designed to help the decoder translate semantic roles in the correct order, as well as generating complete predicate-argument structures. We observed that using a semantic language model can significantly improve the translations, and help the decoder to generate the semantic roles in the correct order. Adding translation rules with complete semantic structures also improved our MT system. We experimented with using source side *complete semantic rules*, as well as using rules for both the source and the target sides. Both of our experiments showed improvements over the baseline, and as expected, the second one had a higher improvement.

Acknowledgments

Partially funded by NSF grant IIS-0910611.

References

Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic roles for string to tree machine translation. In *Association for Computational Linguistics (ACL-13) short paper*.

Tagyoung Chung, Licheng Fang, and Daniel Gildea. 2011. Issues concerning decoding with synchronous context-free grammar. In *Proceedings of the ACL 2011 Conference Short Papers*, Portland, Oregon. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. 2007. Learning Bilingual Semantic Frames: Shallow Semantic Parsing vs. Semantic Role Projection. In *TMI-2007: Proceedings of the 11 th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde, Sweden.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of NAACL-04*, pages 273–280, Boston, MA.

Mark Hopkins and Greg Langmead. 2010. SCFG decoding without binarization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 646–655, Cambridge, MA, October.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395, Barcelona, Spain, July.

Junhui Li, Philip Resnik, and Hal Daumé III. 2013. Modeling syntactic and semantic structures in hierarchical phrase-based translation. In *HLT-NAACL*, pages 540–549.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *COLING-10*, Beijing.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL-03*, pages 160–167, Sapporo, Japan.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha. Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon ConferenceGenLex-09*, Pisa, Italy, Sept.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL-02*, pages 311–318, Philadelphia, PA.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, volume 2, pages 901–904.

Dekai Wu and Pascale Fung. 2009. Semantic roles for smt: A hybrid two-pass model. In *Proceedings of the HLT-NAACL 2009: Short Papers*, Boulder, Colorado.

Shumin Wu and Martha Palmer. 2011. Semantic mapping using automatic word alignment and semantic role labeling. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, pages 21–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for smt. In *ACL (1)*, pages 902–911.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of ACL-02*, pages 303–310, Philadelphia, PA.