Corpus Expansion for Statistical Machine Translation with Semantic Role Label Substitution Rules

Qin Gao and **Stephan Vogel**
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{qing, stephan.vogel}@cs.cmu.edu

Abstract

We present an approach of expanding parallel corpora for machine translation. By utilizing Semantic role labeling (SRL) on one side of the language pair, we extract SRL substitution rules from existing parallel corpus. The rules are then used for generating new sentence pairs. An SVM classifier is built to filter the generated sentence pairs. The filtered corpus is used for training phrase-based translation models, which can be used directly in translation tasks or combined with baseline models. Experimental results on Chinese-English machine translation tasks show an average improvement of 0.45 BLEU and 1.22 TER points across 5 different NIST test sets.

1 Introduction

Statistical machine translation (SMT) relies on parallel corpus. Aside from collecting parallel corpus, we have seen interesting research on automatically generating corpus from existing resources. Typical examples are paraphrasing using bilingual (Callison-Burch et al., 2006) or monolingual (Quirk et al., 2004) data. In this paper, we propose a different methodology of generating additional parallel corpus. The basic idea of paraphrasing is to find alternative ways that convey the **same information**. In contrast, we propose to build new parallel sentences that convey **different information**, yet retain correct grammatical and semantic structures.

The basic idea of the proposed method is to substitute source and target phrase pairs in a sentence pair with phrase pairs from other sentences. The problem is how to identify where a substitution should happen and which phrase pairs are valid candidates for the substitution. While syntactical constraints have been proven to helpful in identifying good paraphrases (Callison-Burch, 2008), it is insufficient in our task because it cannot properly filter the candidates for the replacement. If we allow all the NPs to be replaced with other NPs, each sentence pair can generate huge number of new sentences. Instead, we resort to Semantic Role Labeling (Palmer et al., 2005) to provide more lexicalized and semantic constraints to select the candidates. The method only requires running SRL labeling on either side of the language pair, and that enables applications on low resource languages. Even with the SRL constraints, the generated corpus may still be large and noisy. Hence, we apply an additional filtering stage on the generated corpus. We used an SVM classifier with features derived from standard phrase based translation models and bilingual language models to identify high quality sentence pairs, and use these sentence pairs in the SMT training. In the remaining part of the paper, we introduce the approach and present experimental results on Chinese-to-English translation tasks, which showed improvements across 5 NIST test sets.

2 The Proposed Approach

The objective of the method is to generate new syntactically and semantically well-formed parallel sentences from existing corpus. To achieve this, we first collect a set of rules as the candidates for the substitution. We also need to know where we should put in the replacements and whether the resulting sentence pairs are grammatical.

First, standard word alignment and phrase extraction are performed on existing corpus. Afterwards, we apply an SRL labeler on either the source or target language, whichever has a better SRL labeler. Third, we extract SRL substitution rules (SSRs) from the corpus. The rules carry information of semantic frames, semantic roles, and corresponding

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 294–298,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

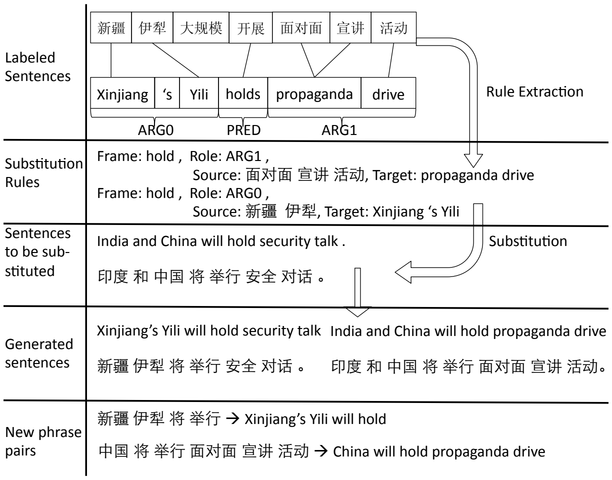
Labeled Sentences
新疆
伊犁
大规模
开展
面对面
宣讲
活动
Xinjiang
‘s
Yili
holds
propaganda
drive
ARG0
PRED
ARG1
Rule Extraction
Substitution Rules
Frame: hold , Role: ARG1 ,
Source: 面对面 宣讲 活动, Target: propaganda drive
Frame: hold , Role: ARG0 ,
Source: 新疆 伊犁, Target: Xinjiang ‘s Yili
Sentences to be substituted
India and China will hold security talk .
印度 和 中国 将 举行 安全 对话 。
Substitution
Generated sentences
Xinjiang’s Yili will hold security talk
India and China will hold propaganda drive
新疆 伊犁 将 举行 安全 对话 。
印度 和 中国 将 举行 面对面 宣讲 活动。
New phrase pairs
新疆 伊犁 将 举行 → Xinjiang’s Yili will hold
中国 将 举行 面对面 宣讲 活动 → China will hold propaganda drive

Figure 1: Examples of extracting SSR and applying them on new sentences. New phrases that will otherwise not be included in the phrase table are shown on the bottom.

source and target phrases. Fourth, we replace phrase pairs in existing sentences with the SSR if they have the same semantic frames and semantic roles.

The newly generated sentence pairs will pass through a classifier to determine whether they are acceptable parallel sentences. And, finally, we train MT system using the new corpus. The resulting phrase table can either be used directly in translation tasks or be interpolated with baseline phrase tables.

3 SRL Substitution Rules

Swapping phrase pairs that serve as the same semantic role of the same semantic frame can provide more combinations of words and phrases. Figure 1 shows an example. The phrase pair “新疆 伊犁 将 举行 → Xinjiang’s Yili will hold” would not be observed in the original corpus without substitution. In this paper, we call *a tuple of semantic frame and semantic role* a *semantic signature*. Two phrase pairs with the same semantic signature are considered valid substitutions of each other.

The extraction of SSRs is similar to the well-known phrase extraction algorithm (Och and Ney, 2004). The criteria of a phrase pair to be included in the SSR set are[1]:

- The phrase on side A must cover a whole semantic role constituent, and it must not contain words in any other semantic role constituent of the same frame.
- The phrase on side B must not contain words that link to words not in the phrase on side A.
- Both of the two boundary words on side B phrases must have at least one link to a word of the phrases on side A. The boundary words on side A phrases can be unaligned only if they are inside the semantic role constituent.

Utilizing these rules, we can perform the sentence generation process. For each semantic structure of each sentence,[2] we determine the phrase pair to be replaced by the same criteria as mention above, and search for suitable SSRs with the same semantic signature. Finally, we replace the original phrases with the source and target side phrases given by the SSRs. Notice that for each new sentence generated, we allow for application of only one substitution.

Although the idea is straightforward, we face two problems in practice. First, for frequent semantic frames, the number of substitution candidates can be very large. It will generate many new sentence pairs, and can easily exceed the capacity of our system. To deal with the problem, we pre-filter the SSRs so that each semantic signature is associated with no more than 100 SSRs. As we can see from the criteria for extracting SSRs, all the entries in the SSR rule set satisfies the commonly used phrase extraction heuristics. Therefore, the set of SSRs is a subset of the phrase table. Because of this, We use the features in the phrase table to sort the rules, and keep 100 rules with highest the arithmetic mean of the feature values.

The second problem is the phrase boundaries are often inaccurate. To handle this problem, we use a simple “glue” algorithm during the substitution. If the inserted phrase has a prefix or suffix sub-phrase that is the same as the suffix or prefix of the adjacent parts of the original sentence, then the duplication will be removed.

4 Classification of Generated Sentences

We can expect the generated corpus be noisy, and needs to be filtered. In this paper we use an SVM classifier to perform this task. First we label a set of

[1] We call the language which has SRL labels side A, and the other language side B.

[2] One sentence can have multiple semantic structures.

sentence pairs [3] randomly sampled from the generated data. We ask the following questions:

1. Are the two sentences grammatical, especially on the boundaries of substituted phrase pairs?
2. Are the two sentences still parallel?

If both questions have positive answers, we label the sentence pair as positive. We can then use the labels together with the features to train the classifier. It is worth mentioning that when we say "grammatical", we do not care about the validity of the actual meaning of the sentence.

The set of SSR is a subset of the phrase table. Therefore, the features in the phrase table can be used as features. It includes the bidirectional phrase and lexicon translation probabilities.

In addition, we use the language model features. The language model score of the whole sentence is useless because it is dominated by words not affected by the substitution. Therefore, we only consider n-grams that are affected by the substitution. I.e. only the boundary words are taken into account. Given an n-gram language model, we only calculate the scores in windows with the size $2n-2$, centered on the boundary of the substituted phrases. In other words, $n-1$ words before and after the boundaries will be included in the calculation.

Finally, there are two additional features: the probability of observing the source/target phrase given the semantic signature. They can be calculated by counting the frequencies of source/target phrases and the semantic signature in extracted rules.

As we have abundant sentence pairs generated, we prefer to apply a more harsh filtering, keeping only the best candidates. Therefore, when training the SVM model, we intentionally increase the cost of false positive errors, so as to maximize the precision rate of positive decisions and reduce possible contamination. In an experiment, we used 900 of the 1000 labeled sentence pairs as the training set, and the remaining 100 (41 positive and 59 negative samples) sentence pairs as the test set. By setting the cost of false positive errors to 1.33, we classified 20 of 41 positive samples correctly, and only 3 of the 59 negative samples are classified as positive.

Corpus	Sents.	Words		Avg. Sent. Len	
		Ch	En	Ch	En
Baseline	387K	11.2M	14.7M	28.95	38.19
Before-Filter	29.6M	970M	1.30B	32.75	44.08
After-Filter	7.2M	239M	306M	32.92	42.16
GALE	8.7M	237M	270M	27.00	30.69

Table 1: Statistics of generated corpus.

5 Utilizing the Generated Corpus

With the generated corpus, we perform training and generate a new phrase table. There are many ways of utilizing the new phrase table; the simplest way is to use it directly for translation tasks. However, the new phrase table may be noisier than the original one. To solve this, we interpolate the new phrase table with the baseline phrase table. If a phrase pair is only observed in the baseline phrase table, we keep it intact in the interpolated phrase table. If a phrase pair is observed only in the new phrase table, we discount all the feature values by a factor of 2. And if the phrase pair is in both of the phrase tables, the feature values will be the arithmetic mean of the corresponding values in the two phrase tables.

We also noticed that the new corpus may have very different distribution of words comparing to the baseline corpus. The word alignment process using generative models is more likely to be affected by the radical change of distributions. Therefore, we also experimented with force aligning the generated corpus with the word alignment models trained baseline corpus before building the phrase table.

6 Experiments

We performed experiments on Chinese to English MT tasks with the proposed approach. The baseline system is trained on the FBIS corpus, the statistics of the corpus is shown in Table 1. We adopted the ASSERT English SRL labeler (Pradhan et al., 2004), which was trained on PropBank data using SVM classifier. The labeler reports 81.87% precision and 73.21% recall rate on CoNLL-2005 shared task on SRL. We aligned the parallel sentences with MGIZA(Gao and Vogel, 2008), and performed experiments with the Moses toolkit (Koehn et al, 2007).

The rule extraction algorithm produces 1.3 mil-

[3] We manually labeled 1000 sentence pairs

	BLEU scores				
	mt02	mt03	mt04	mt05	mt08	avg
BL	32.02	29.75	33.12	29.83	**24.15**	n/a
GS	31.09	29.39	32.86	29.29	23.57	-0.53
IT	32.41	**30.70**	**33.91**	**30.30**	23.80	**+0.45**
GA	**32.57**	30.13	33.50	30.42	23.87	+0.32
IA	32.20	29.62	33.08	29.37	24.09	-0.10
LS	32.52	*31.67*	33.36	*31.58*	*24.81*	*+1.01*

	TER scores for Full FBIS Corpus					
	mt02	mt03	mt04	mt05	mt08	avg
BL	68.94	70.21	66.67	70.35	69.33	n/a
GS	69.97	70.22	66.74	70.32	69.96	+0.34
IT	68.04	68.52	65.19	68.83	68.80	-1.22
GA	**67.12**	**68.38**	**64.75**	**67.90**	**68.37**	**-1.80**
IA	68.54	69.88	66.07	70.08	68.98	-0.39
LS	68.15	68.56	66.01	68.71	69.37	-0.94

Table 2: Experiment results on Chinese-English translation tasks, the abbreviations for systems are as follows: BL: Baseline system, GS: System trained with only generated sentence pairs, IT: Interpolated phrase table with GS and BL,. GA and IA are GS and IT systems trained with baseline word alignment models accordingly. LS is the GALE system with 8.7M sentence pairs.

lion SSRs. As we can observe in Table 1, we generated 29.6 million sentences from the 387K sentence pairs, and by using the SVM-based classifier, we filter the corpus down to 7.2 million. We also observed that the average sentence length increases by 15% in the generated corpus. That is because longer sentences have more slots for substitution. Therefore, they have more occurrences in the generated corpus.

We used the NIST MT06 test set for tuning, and experimented with 5 test sets, including MT02, 03, 04, 05, 08. Table 2 shows the BLEU and TER scores of the experiments. As we can see in the results, by using only the generated sentence pairs, the performance of the system drops. However the interpolated phrase tables outperform the baseline. On average, the improvements on all the 5 test sets are 0.45 on BLEU score and -1.22 on TER when using the interpolated phrase table. We do observe MT08 drops on BLEU scores; however, the TER scores are consistently improved across all the test sets. When using baseline alignment model, we observe a quite different phenomenon. In this case, interpolating the phrase tables no longer show improvements. However, using the generated corpus alone achieves -1.80 on average TER. An explanation is that using identical alignment model makes the phrases extracted from the baseline and generated corpus similar, which undermines the idea of interpolating two phrase tables. As shown in Table 3, it generates less new source phrases and 10% more phrase pairs that overlaps with the baseline phrase table. For comparison, we also provide scores from a system that uses the training data for GALE project, which has 8.7M sentence pairs[4]. In Table 3 we observe that the large GALE system yields better BLEU results while the IT or GA systems have even better TER scores than the GALE system. The expanded corpus performs almost as well as the GALE system even though the large system has a phrase table that is four time larger.

	PT size	C.P.	D.S.	N.S.	T/S	A.L.
BL	30.0M	100%	12.5M	0	2.40	1.46
GS	78.6M	46%	35.4M	28.2M	2.22	1.49
IT	94.6M	100%	40.7M	28.2M	2.32	1.56
GA	79.4M	56%	35.5M	27.7M	2.24	1.54
IA	92.7M	100%	40.2M	27.7M	2.30	1.52
LS	352M	55%	147.2M	142.7M	2.40	1.63

Table 3: Statistics of phrase tables and translation outputs, including the phrase tables (PT) size, the coverage of the BL phrase table entries (C.P.), the number of source phrases (D.S.), the number of new source phrases comparing to BL system (N.S.), the average number of alternative translations of each source phrase (T/S) and the average source phrase length in the output (A.L.)

The statistics of the phrase tables and translation outputs are listed in Table 3. As we can see, the generated sentence introduces a large number of new source phrases and the average lengths of matching source phrases of all the systems are longer than the baseline, which could be an evidence for our claim that the proposed approach can generate more high quality sentences and phrase pairs that have not been observed in the original corpus.

7 Conclusion

In this paper we explore a novel way of generating new parallel corpus from existing SRL labeled corpus. By extracting SRL substitution rules (SSRs) we generate a large set of sentence pairs, and by applying an SVM-based classifier we can filter the corpus,

[4] FBIS corpus is included in the GALE dataset

keeping only grammatical sentence pairs. By interpolating the phrase table with the baseline phrase table, we observed improvement on Chinese-English machine translation tasks and the performance is comparable to system trained with larger manually collected parallel corpus. While our experiments were performed on Chinese-English, the approach is more useful for low resource languages. The advantage of the proposed method is that we only need the SRL labels on either side of the language pair, and we can choose the one with a better SRL labeler.

The features we used in the paper are still primitive, which results in a classifier radically tuned against false positive rate. This can be improved by designing more informative features.

Since the method will only introduce new phrases across the phrase boundaries of phrases in existing phrase table, it is desirable to be integrated with other paraphrasing approaches to further increase the coverage of the generated corpus.

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30:417–449, December.

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

Sameer S. Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2004)*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.