Using Parallel Propbanks to enhance Word-alignments

Jinho D. Choi
Dept. of Computer Science
Univ. of Colorado at Boulder
choijd@colorado.edu

Martha Palmer
Dept. of Linguistics
Univ. of Colorado at Boulder
mpalmer@colorado.edu

Nianwen Xue
Dept. of Computer Science
Brandeis University
xuen@brandeis.edu

Abstract

This short paper describes the use of the linguistic annotation available in parallel PropBanks (Chinese and English) for the enhancement of automatically derived word alignments. Specifically, we suggest ways to refine and expand word alignments for verb-predicates by using predicate-argument structures. Evaluations demonstrate improved alignment accuracies that vary by corpus type.

1 Introduction

Since verbs tend to be the roots of dependency relations in a sentence (Palmer et al., 2005), when it comes down to translations, finding correct mappings between verbs in a source and a target language is very important. Many machine translation systems (Fraser and Marcu, 2007) use word-alignment tools such as GIZA++ (Och and Ney, 2003) to retrieve word mappings between a source and a target language. Although GIZA++ gives well-structured alignments, it has limitations in several ways. First, it is hard to verify if alignments generated by GIZA++ are correct. Second, GIZA++ may not find alignments for low-frequent words. Third, GIZA++ does not account for any semantic information.

In this paper, we suggest a couple of ways to enhance word-alignments for predicating expressions such as verbs[1]. We restricted the source and the target language to Chinese and English, respectively. The goal is to use the linguistic annotation available in parallel PropBanks (Xue and Palmer, 2009) to refine and expand automatic word-alignments. First, we check if the alignment for each Chinese predicate, generated by GIZA++, is also a predicate in English (Section 3). If it is, we verify if the alignment is correct by matching their arguments (Section 4.1). If it is not, we find an English predicate that has the maximum argument matching with the Chinese predicate (Section 4.2). Finally, we evaluate the potential of the enhanced word-alignments for providing a significant improvement over the GIZA++ baseline.

2 Parallel Corpus

We used the 'English Chinese Translation Treebank' (ECTB), a parallel English-Chinese corpus. In addition to the treebank syntactic structure, the corpus has also been annotated with semantic role labels in the standard PropBank style of Arg0, Arg1, etc., based on verb specific frame file definitions (Xue and Palmer, 2009). The corpus is divided into two parts: the Xinhua Chinese newswire with literal English translations (4,363 parallel sentences) and the Sinorama Chinese news magazine with non-literal English translations (12,600 parallel sentences). We experimented with the two parts separately to see how literal and non-literal translations affect word-alignments.

3 Predicate Matching

For preprocessing, we ran GIZA++ on ECTB to get word-alignments between Chinese and English. Then, for each Chinese predicate, we checked if it is aligned to an English predicate by using the gold-standard parallel Propbanks. Table 1 shows how many Chinese predicates were aligned to what kind of English words.

Only (45.3%-Xinhua, 19.1%-Sinorama) of Chinese predicates were aligned to words that are predicates in English. It is true that not all Chinese verbs are supposed to be translated to verbs in English, but that does not account for the numbers in Table 1. We therefore assume that there are opportunities to enhance word-alignments for Chinese and English predicates.

[1] Throughout the paper, all predicates refer to verbs.

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 121–124,
Suntec, Singapore, 6-7 August 2009.

Alignment	Xinhua	Sinorama
Ch.pred → En.pred	5,842	7,643
Ch.pred → En.be	386	1,229
Ch.pred → En.else	2,489	8,726
Ch.pred → En.none	4,178	22,488
Total	12,895	40,086

Table 1: Results of predicate matching (Ch: Chinese, En: English, pred: predicates, be: be-verbs, else: non-verbs, none: no word). The numbers indicate the amount of verb-tokens, not verb-types.

4 Argument Matching

For Chinese predicates aligned to English predicates, we can verify the alignments by 'Top-down argument matching': given Chinese and English predicates that are aligned, check if their arguments are also aligned (arguments are found from parallel Propbanks). The intuition is that if the predicates are correctly aligned across the languages, their arguments should be aligned as well.

For Chinese predicates not aligned to any English words, we can find their potential English alignments by 'Bottom-up argument matching': given a set of arguments for a such Chinese predicate, find some English predicate whose set of arguments has the most words aligned to words in the Chinese arguments. If the words in the arguments are mostly aligned (above a certain threshold) across the languages, we suspect that the predicates should be aligned as well.

4.1 Top-down Argument Matching (T-D)

Given a Chinese predicate p_c aligned to an English predicate p_e, let S_c and S_e be a set of arguments for p_c and p_e, respectively. For each $ca_i \in S_c$, we match it with some $ea_j \in S_e$ that has the most words aligned to words in ca_i. If such ea_j exists, we count the number of aligned words, say $|ca_i \cap ea_j|$; otherwise, the count is 0. Once the matchings are done, we average the proportions of the counts and if the average is above a certain threshold, we consider the alignment is correct.

Let us look at the example in Table 2. After the preprocessing, a Chinese predicate '设立' is aligned to an English predicate 'set up' by GIZA++. '设立' has two arguments, Ch.Arg0 and Ch.Arg1, retrieved from the Chinese Propbank. For each Chinese argument, we search for some argument of 'set' (from the English Propbank) that has the most words aligned. For instance, words in Ch.Arg0, '这些 城市', are aligned to 'those municipalities' by GIZA++ so Ch.Arg0 finds En.Arg0 as the one maximizes word-interscetions (similar for Ch.Arg1 and En.Arg1). In this case, the argument matchings for all pairs of arguments are 100%, so we consider the alignment is correct.

– Chinese Sentence –
: 同时 还 批准 这些 城市 设立 十四 个 边境 经济 合作区
- **Predicate**: 设立.01 → set up
- **Ch.Arg0**: 这些 城市 → those municipalities
- **Ch.Arg1**: 十四 个 边境 经济 合作区
→ fourteen border economic cooperation zones

– English Sentence –
: At the same time it also sanctioned those municipalities to set up fourteen border economic cooperation zones
- **Predicate**: set.03 (set up)
- **En.Arg0**: those municipalities
- **En.Arg1**: fourteen border economic cooperation zones

Table 2: Parallel sentences labelled with their semantic roles

Table 3 shows the average argument matching scores for all pairs of Chinese and English predicates. For each pair of predicates, 'macro-average' measures the proportion of word-intersections for each pair of Chinese and English arguments (with the most words aligned) and averages the proportions whereas 'micro-average' counts word-intersections for all pairs of arguments (each pair with the most words aligned) and divides it by the total number of words in Chinese arguments.

- S_c = a set of Chinese arguments, $ca_i \in S_c$
- S_e = a set of English arguments, $ea_j \in S_e$
- Macro average argument matching score
$$= \frac{1}{|S_c|} \sum_{\forall ca_i} \left(\frac{argmax(|ca_i \cap ea_j|)}{|ca_i|} \right)$$
- Micro average argument matching score
$$= \frac{\sum_{\forall ca_i} argmax(|ca_i \cap ea_j|)}{\sum_{\forall ca_i} |ca_i|}$$

	Xinhua	Sinorama
Macro Avg.	80.55%	53.56%
Micro Avg.	83.91%	52.62%

Table 3: Average argument matching scores for top-down argument matching

It is not surprising that Xinhua's scores are higher because the English sentences in Xinhua are more literally translated than ones in Sinorama so that it is easier to find correct alignments in Xinhua.

4.2 Bottom-Up Argument Matching (B-U)

A large portion of Chinese predicates are aligned to no English words. For such Chinese predicate, say p_c, we check to see if there exists an English predicate within the parallel sentence, say p_e, that is not aligned to any Chinese word and gives the maximum micro-average score (Section 4.1) compare to all other predicates in the English sentence. If the micro-average score is above a certain threshold, we align p_c to p_e.

The thresholds we used are 0.7 and 0.8. Thresholds below 0.7 assumes too many alignments that are incorrect and ones above 0.8 assumes too few alignments to be useful. Table 4 shows the average argument matching scores for alignments found by bottom-up argument matching.

	Xinhua		**Sinorama**	
Thresh.	**0.7**	**0.8**	**0.7**	**0.8**
Macro	80.74	83.99	77.70	82.86
Micro	82.63	86.46	79.45	85.07

Table 4: Average argument matching scores in percentile for bottom-up argument matching

5 Evaluations

Evaluations are done by a Chinese-English bilingual. We used a different English-Chinese parallel corpus for evaluations. There are 100 parallel sentences, 365 Chinese verb-tokens, and 273 Chinese verb-types in the corpus. We tested word-alignments, refined and expanded by our approaches, on verb-types rather than verb-tokens to avoid over-emphasizing multiple appearances of a single type. Furthermore, we tested word-alignments from Xinhua and Sinorama separately to see how literal and non-literal translations affect the outcomes.

5.1 Refining word-alignment

We used three kinds of measurements for comparisons: term coverage, term expansion, and alignment accuracy. 'Term coverage' shows how many source terms (Chinese verb-types) are covered by word-alignments found in each corpus. Out of 273 Chinese verb-types in the test corpus, (79-Xinhua, 129-Sinorama) were covered by word-alignments generated by GIZA++. 'Term expansion' shows how many target terms (English verb-types) are suggested for each of the covered source terms. There are on average (1.77-Xinhua, 2.29-Sinorama) English verb-types suggested for each covered Chinese verb-type. 'Alignment accuracy' shows how many of the suggested target terms are correct. Among the suggested English verb-types, (83.35%-Xinhua, 57.76%-Sinorama) were correct on average.

The goal is to improve the alignment accuracy with minimum reduction of the term coverage and expansion. To accomplish the goal, we set a threshold for the T-D's macro-average score: for Chinese predicates aligned to English predicates, we kept only alignments whose macro-average scores meet or exceed a certain threshold. The thresholds we chose are 0.4 and 0.5; lower thresholds did not have much effect and higher thresholds threw out too many alignments. Table 5 shows the results of three measurements with respect to the thresholds (Note that all these alignments were generated by GIZA++).

	Xinhua			**Sinorama**		
TH	**TC**	**ATE**	**AAA**	**TC**	**ATE**	**AAA**
0.0	**79**	**1.77**	83.35	**129**	**2.29**	57.76
0.4	76	1.72	83.54	93	1.8	65.88
0.5	76	1.68	**83.71**	62	1.58	**78.09**

Table 5: Results for alignment refinement (TH: threshold, TC: term coverage, ATE: average term expansion, AAA: average alignment accuracy in percentage). The highest score for each measurement is marked as bold.

As you can see, thresholds did not have much effect on alignments found in Xinhua. This is understandable because the translations in Xinhua are so literal that it was relatively easy for GIZA++ to find correct alignments; in other words, the alignments generated by GIZA++ were already very accurate. However, for alignments found in Sinorama, the average alignment accuracy increases radically as the threshold increases. This implies that it is possible to refine word-alignments found in a corpus containing many non-literal translations by using T-D.

Notice that the term coverage for Sinorama decreases as the threshold increases. Considering how much improvement it made for the average alignment accuracy, we suspect that it filtered out mostly ones that were incorrect alignments.

5.2 Expanding word-alignment

We used B-U to expand word-alignments for Chinese predicates aligned to no English words. We decided not to expand alignments for Chinese predicates aligned to non-verb English words because GIZA++ generated alignments are more accurate than ones found by B-U in general.

There are (22-Xinhua, 20-Sinorama) additional verb-types covered by the expanded-alignments. Note that these alignments are already filtered by the micro-average score (Section 4.2). To refine the alignments even more, we set a threshold on the macro-average score as well. The thresholds we used for the macro-average score are 0.6 and 0.7. Table 6 shows the results of the expanded-alignments found in Xinhua and Sinorama.

	Mac - 0.7			**Mac - 0.8**		
	TC	**ATE**	**AAA**	**TC**	**ATE**	**AAA**
Mic	**Xinhua**					
0.0	**22**	**4.27**	50.38	**20**	3.35	57.50
0.6	21	3.9	54.76	18	**3.39**	**63.89**
0.7	19	3.47	**55.26**	17	3.12	61.76
Mic	**Sinorama**					
0.0	**37**	**3.59**	**18.01**	**29**	**3.14**	**14.95**
0.6	31	3.06	15.11	27	2.93	14.46
0.7	21	2.81	11.99	25	2.6	11.82

Table 6: Results for expanded-alignments found in Xinhua and Sinorama (Mac: threshold on macro-average score, Mic: threshold on micro-average score)

The average alignment accuracy for Xinhua is encouraging; it shows that B-U can expand word-alignments for a corpus with literal translations. The average alignment accuracy for Sinorama is surprisingly low; it shows that B-U cannot function effectively given non-literal translations.

6 Summary and Future Works

We have demonstrated the potential for using parallel Propbanks to improve statistical verb translations from Chinese to English. Our B-U approach shows promise for expanding the term-coverage of GIZA++ alignments that are based on literal translations. In contrast, our T-D is most effective with non-literal translations for verifying the alignment accuracy, which has been proven difficult for GIZA++.

This is still a preliminary work but in the future, we will try to enhance word-alignments by using automatically labelled Propbanks, Nombanks (Meyers et al., 2004), Named-entity tagging, and test the enhancement on bigger corpora. Furthermore, we will also evaluate the integration of our enhanced alignments with statistical machine translation systems.

Acknowledgments

Special thanks to Daniel Gildea, Ding Liu (University of Rochester) who provided word-alignments, Wei Wang (Information Sciences Institute at University of Southern California) who provided the test-corpus, and Hua Zhong (University of Colorado at Boulder) who performed the evaluations. We gratefully acknowledge the support of the National Science Foundation Grants IIS-0325646, Domain Independent Semantic Parsing, CISE-CRI-0551615, Towards a Comprehensive Linguistic Annotation, and a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, subcontract from BBN, Inc. Any contents expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Natural Language Engineering*, 15(1):143–172.