HLT-NAACL 2003 Workshop: Building and Using Parallel Texts
Data Driven Machine Translation and Beyond , pp. 15-18
Edmonton, May-June 2003

Word Alignment Based on Bilingual Bracketing

Bing Zhao
Language Technologies Institute
Carnegie Mellon University
bzhao@cs.cmu.edu

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
vogel+@cs.cmu.edu

Abstract

In this paper, an improved word alignment based on bilingual bracketing is described. The explored approaches include using Model-1 conditional probability, a boosting strategy for lexicon probabilities based on importance sampling, applying Parts of Speech to discriminate English words and incorporating information of English base noun phrase. The results of the shared task on French-English, Romanian-English and Chinese-English word alignments are presented and discussed.

1 Introduction

Bilingual parsing based word alignment is promising but still difficult. The goal is to extract structure information from parallel sentences, and thereby improve word/phrase alignment via bilingual constraint transfer. This approach can be generalized to the automatic acquisition of a translation lexicon and phrase translations esp. for languages for which resources are relatively scarce compared with English.

The parallel sentences in building Statistical Machine Translation (SMT) systems are mostly unrestricted text where full parsing often fails, and robustness with respect to the inherent noise of the parallel data is important. Bilingual Bracketing [Wu 1997] is one of the bilingual shallow parsing approaches studied for Chinese-English word alignment. It uses a translation lexicon within a probabilistic context free grammar (PCFG) as a generative model to analyze the parallel sentences with weak order constraints. This provides a framework to incorporate knowledge from the English side such as POS, phrase structure and potentially more detailed parsing results.

In this paper, we use a simplified bilingual bracketing grammar together with a statistical translation lexicon such as the Model-1 lexicon [Brown 1993] to do the bilingual bracketing. A boosting strategy is studied and applied to the statistical lexicon training. English POS and Base Noun Phrase (NP) detection are used to further improve the alignment performance. Word alignments and phrase alignments are extracted from the parsing results as post processing. The settings of different translation lexicons within the bilingual bracketing framework are studied and experiments on word-alignment are carried out on Chinese-English, French-English, and Romanian-English language pairs.

The paper is structured as follows: in section 2, the simplified bilingual bracketing used in our system is described; in section 3, the boosting strategy based on importance sampling for IBM Model-1 lexicon is introduced; in section 4, English POS and English Base Noun Phrase are used to constrain the alignments ; in section 5, the experimental results are shown; summary and conclusions are given in section 6.

2 Bilingual Bracketing

In [Wu 1997], the Bilingual Bracketing PCFG was introduced, which can be simplified as the following production rules:

$$A \rightarrow [AA] \quad (1)$$
$$A \rightarrow < AA > \quad (2)$$
$$A \rightarrow f/e \quad (3)$$
$$A \rightarrow f/null \quad (4)$$
$$A \rightarrow null/e \quad (5)$$

Where *f* and *e* are words in the target vocabulary V_f and source vocabulary V_e respectively. A is the alignment of texts. There are two operators for bracketing: direct bracketing denoted by [ ], and inverse bracketing, denoted by <>. The A-productions are divided into two classes: syntactic {(1),(2)}and lexical rules {(3),(4),(5)}. Each A-production rule has a probability.

In our algorithm, we use the same PCFG. However, instead of estimating the probabilities for the production rules via EM as described in [Wu 1997], we assign the probabilities to the rules using the Model-1 statistical translation lexicon [Brown et al. 1993].

Because the syntactic A-production rules do not compete with the lexical rules, we can set them some default values. Also we make no assumptions which bracketing direction is more likely to occur, thus the probabilities for [ ] and <> are set to be equal. As for the lexical rules, we experimented with the conditional probabilities $p(e|f)$, $p(f|e)$ and the interpolation of $p(f|e, e_{pos})$ and $p(f|e)$ (described in section 4.1). As for these probabilities of aligning a word to the null word or to unknown words, they are set to be 1e-7, which is the default small value used in training Model-1.

The word alignment can then be done via maximizing the likelihood of matched words subject to the bracketing grammar using dynamic programming.

The result of the parsing gives bracketing for both input sentences as well as bracket alignments indicating the corresponding brackets between the sentence pairs. The bracket alignment includes a word alignment as a by-product. One example for French-English (the test set sentence pair #18) is shown as below:

[[it_1 is_2] [$quite_3$ [$understandable_4$ $_{.5}$]]]
[[ce_1 est_2] [$tout_3$ [[‘$_4$ [$fait_5$ $comprihensible_6$]] $_{.7}$]]]

[[it_1/ce_1 is_2/est_2] [$quite_3/tout_3$ [[e/‘$_4$ [$e/fait_5$ $understandable_4/comprihensible_6$]] $_{.5}/_{.7}$]]]

3 Boosting Strategy of Model-1 Lexicon

The probabilities for the lexical rules are Model-1 conditional probabilities $p(f|e)$, which can be estimated using available toolkits such as [Franz 2000].

This strategy is a three-pass training of Model-1, which was shown to be effective in our Chinese-English alignment experiments. The first two passes are carried out to get Viterbi word alignments based on Model-1's parameters in both directions: from source to target and then vice versa. An intersection of the two Viterbi word alignments is then calculated. The highly frequent word-pairs in the intersection set are considered to be important samples supporting the alignment of that word-pair. This approach, which is similar to importance sampling, can be summarized as follows:

Denote a sample as a co-occurred word-pair as $x = (e_i, f_j)$ with its observed frequency: $C(x) = freq(e_i, f_j)$; Denote $I(x) = freq(e_i, f_j)$ as the frequency of that word-pair x observed in the intersection of the two Viterbi alignments.

- Build $I(x) = freq(e_i, f_j)$ from the intersection of alignments in two directions.
- Generate $x = (e_i, f_j)$ and its $C(x) = freq(e_i, f_j)$ observed from a given parallel corpus;
- Generate random variable u from uniform [0,1] distribution independent of x;
- If $\frac{I(x)}{M \cdot C(x)} \geq u$, then accept x, where M is a finite known constant $M > 0$;
- Re-weight sample x: $C_b(x) = C(x)*(1+\varepsilon), \varepsilon > 0)$

The modified counts (weighted samples) are re-normalized to get a proper probability distribution, which is used in the next iteration of EM training. The constant M is a threshold to remove the potential noise from the intersection set. M's value is related to the size of the training corpus, the larger its size, the larger M should be. ε is chosen as a small positive value. The overall idea is to collect those word-pairs which are reliable and give an additional pseudo count to them.

4 Incorporating English Grammatical Constraints

There are several POS taggers, base noun phrase detectors and parsers available for English. Both the shallow and full parsing information of English sentences can be used as constraints in Bilingual Bracketing. Here, we explored utilizing English POS and English base noun phrase boundaries.

4.1 Incorporating English POS

The correctly aligned words from two languages are very likely to have the same POS. For example, a Chinese noun is very likely to be aligned with a English noun. While the English POS tagging is often reliable and accurate, the POS tagging for other languages is usually not easily acquired nor accurate enough. Modelling only the English POS in word alignment is usually a practical way.

Given POS information for only the English side, we can discriminate English words and thus disambiguate the translation lexicon. We tagged each English word in the parallel corpus, so that each English word is associated with its POS denoted as e_{pos}. The English word and its POS were concatenated into one pseudo word. For example: beginning/NN and beginning/VBG are two pseudo words which occurred in our training corpus. Then the Model-1 training was carried out on this concatenated parallel corpus to get estimations of $p(f|e, e_{pos})$.

One potential problem is the estimation of $p(f|e, e_{pos})$. When we concatenated the word with its POS, we implicitly increased the vocabulary size. For example, for French-English training set, the English vocabulary increased from 57703 to 65549. This may not cause a problem when the training data's size is large. But for small

parallel corpora, some correct word-pair's $p(f|e, e_{pos})$ will be underestimated due to the sparse data, and some word-pairs become unknown in $p(f|e, e_{pos})$. So in our system, we actually interpolated $p(f|e, e_{pos})$ with $p(f|e)$ as a mixture model for robustness:

$$P(A \rightarrow f/e|A) = \lambda \cdot P(f|e) + (1-\lambda) \cdot P(f|e, e_{pos}) \quad (6)$$

Where λ can be estimated by EM for this two-mixture model on the training data, or a grid search via cross-validation.

4.2 Incorporating English Base Noun Boundaries

The English sentence is bracketed according to the syntactic A-production rules. This bracketing can break an English noun phrase into separated pieces, which are not in accordance with results from standard base noun phrase detectors. Though the word-alignments may still be correct, but for the phrase level alignment, it is not desired.

One solution is to constrain the syntactic A-production rules to penalize bracketing English noun phrases into separated pieces. The phrase boundaries can be obtained by using a base noun phrase detection toolkit [Ramshaw 1995], and the boundaries are loaded into the bracketing program. During the dynamic programming, before applying a syntactic A-production rule, the program checks if the brackets defined by the syntactic rule violate the noun phrase boundaries. If so, an additional penalty is attached to this rule.

5 Experiments

All the settings described so far are based on our previous experiments on Chinese-English (CE) alignment. These settings are then used directly without any adjustment of the parameters for the French-English (FE) and Romanian-English (RE) word alignment tasks. In this section, we will first describe our experiments on Chinese-English alignment, and then the results for the shared task on French-English and Romanian-English.

For Chinese-English alignment, 365 sentence-pairs are randomly sampled from the Chinese Tree-bank provided by the Linguistic Data Consortium. Three persons manually aligned the word-pairs independently, and the consistent alignments from all of them were used as the reference alignments. There are totally 4094 word-pairs in the reference set. Our way of alignment is very similar to the "SURE" (S) alignment defined in the shared task. The training data we used is 16K parallel sentence-pairs from Hong-Kong news data. The English POS tagger we used is Brill's POS tagger [Brill 1994]. The base noun detector is [Ramshaw 1995]. The alignment is evaluated in terms of precision, recall, F-measure and alignment error rate (AER) defined in the shared task. The results are shown in Table-1:

Table-1. Chinese-English Word-Alignment

CE	precision	recall	F-measure	AER
No-Boost	50.88	58.77	54.54	45.46
Boosted	52.19	60.33	55.96	44.04
+POS	54.77	63.34	58.71	41.29
+NP	55.16	63.75	59.14	40.86

Table-1 shows the effectiveness of using each setting on this small size training data. Here the boosted model gives a noticeable improvement over the baseline. However, our observations on the trial/test data showed very similar results for boosted and non-boosted models, so we present only the non-boosted results(standard Model-1) for the shared task of EF and RE word alignment.

Adding POS further improved the performance significantly. The AER drops from 44.04 to 41.29. Adding additional base noun phrase boundaries did not give as much improvement as we hoped. There is only slight improvement in terms of AER and F-measure. One reason is that noun phrase boundaries is more directly related to phrase alignment than word-alignment. A close examination showed that with wrong phrase-alignment, word-alignment can still be correct. Another reason is that using the noun phrase boundaries this way may not be powerful enough to leverage the English structure information in Bilingual Bracketing. More suitable ways could be bilingual chunk parsing, and refining the bracketing grammar as described in [Wu 1997].

In the shared task experiments, we restricted the training data to sentences upto 60 words. The statistics for the training sets are shown in Table-2. (French/Romanian are source and English is target language).

Table-2.Training Set Statistics

	French-English	Romanian-English
Sent-pairs	1028382	45456
Src Voc	79601	45880
Tgt Voc	57703	26904

There are 447 test sentence pairs for English-French and 248 test sentence pairs for Romanian-English. After the bilingual bracketing, we extracted only the *explicit* word alignment from lexical rules: $A \rightarrow e/f$, where neither e nor f is the null(empty) word. These explicit word alignments are more directly related to the translation quality in our SMT system than the null-word alignments. Also the explicit word alignments is in accordance with the "SURE" (S) alignment defined in the shared tasks. However the Bilingual Bracketing system is not adapted to the "PROBABLE" (P) alignment because of the inherent one-to-one mapping. All the AERs in the following tables are calculated based *solely* on S alignment without any null alignments collected from the bracketing results.

Table-3. Limited Resource French-English

FE	precision	recall	F-measure	AER
$p(f\|e)$	49.85	79.45	61.26	23.87
$p(e\|f)$	51.46	82.42	63.36	20.95
inter	63.03	74.59	68.32	19.26

Table-4. Unlimited Resource French-English

FE	precision	recall	F-measure	AER
$p(f\|e)$	50.21	80.36	61.80	23.07
$p(e\|f)$	51.91	83.26	63.95	19.96
inter	66.34	74.86	70.34	17.77

Table-5. Limited Resource Romanian-English

RE	precision	recall	F-measure	AER
$p(r\|e)$	70.65	55.75	62.32	37.66
$p(e\|r)$	71.39	55.00	62.13	37.87
inter	85.48	48.64	62.01	37.99

Table-6. Unlimited Resource Romanian-English

RE	precision	recall	F-measure	AER
$p(r\|e)$	69.63	54.65	61.24	38.76
$p(e\|r)$	70.36	55.50	62.05	37.95
inter	82.09	48.73	61.15	38.85

For the limited resource task, we trained Model-1 lexicons in both directions: from source to target denoted as $p(f|e)$ and from target to source denoted as $p(e|f)$. These two lexicons are then plugged into the Bilingual Bracketing algorithm separately to get two sets of bilingual bracketing word alignments. The intersection of these two sets of word alignments is then collected. The resulting AERs are shown in Table-3 and Table-5 respectively.

For the unlimited resource task, we again tagged the English sentences and base noun phrase boundaries as mentioned before. Then corresponding Model-1 lexicon was trained and Bilingual Bracketing carried out. Using the same strategies as in the limited resource task, we got the results shown in Table-4 and Table-6.

The table above show that adding English POS and base noun detection gave a consistent improvement for all conditions in the French-to-English alignment. The intersection of the two alignments greatly improves the precision, paired with a reduction in recall, still resulting in an overall improvement in F-measure and AER.

For the Romanian-English alignment the POS tagging and noun phrase boundaries did not help. On the small corpus the increase in vocabulary resulted in addition unknown words in the test sentences which introduces additional alignment errors.

Comparing the results of the French-English and Romanian-English alignment tasks we see a striking difference in precision and recall. Whereas the French-English alignment has a low precision and a high recall its the opposite for the Romanian-English alignment. The cause lays in different styles for the manual alignments. The French-English reference set contains both S and P alignments, whereas the Romanian-English reference set was annotated with only S alignments. As a result, there are on average only 0.5 S alignments per word in the FE reference set, but 1.5 S alignments per word in the RE test set.

6 Summary

In this paper we presented our word alignment system based on bilingual bracketing. We introduced a technique to boost lexical probabilities for more reliable word pairs in the statistical lexicon. In addition, we investigated the effects of using POS and noun phrase detection on the English side of the bilingual corpus as constraints for the alignment. We applied these techniques to the French-English and Romanian-English alignment tasks, and in addition to Chinese-English alignment. For Chinese-English and French-English alignments these additional knowledge sources resulted in improvements in alignment quality. Best results were obtained by using the intersection of the source to target and target to source bilingual bracketing alignments. The results show very different behavior of the alignment system on the French-English and Romanian-English tasks which is due to different characteristics of the manually aligned test data. This indicates that establishing a good golden standard for word alignment evaluation is still an open issue.

References

Brown, P. F. and Della Pietra, S. A. and Della Pietra, V. J. and Mercer, R. L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19-2, pp 263-311.

Erik Brill. 1994. Some advances in rule-based part of speech tagging. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Wa., 1994.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. *Proceedings of ACL-00*, pp. 440-447, Hongkong, China.

Lance Ramshaw and Mitchell Marcus 1995. Text Chunking Using Transformation-Based Learning. *Proceedings of the Third ACL Workshop on Very Large Corpora*, MIT, June, 1995.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3):377-404, Sep. 1997.