Adaptive Parallel Sentences Mining from Web Bilingual News Collection

Bing Zhao, Stephan Vogel
Language Technologies Institute, School of Computer Science, Carnegie Mellon University
{bzhao, vogel+}@cs.cmu.edu

Abstract

In this paper a robust, adaptive approach for mining parallel sentences from a bilingual comparable news collection is described. Sentence length models and lexicon-based models are combined under a maximum likelihood criterion. Specific models are proposed to handle insertions and deletions that are frequent in bilingual data collected from the web. The proposed approach is adaptive, updating the translation lexicon iteratively using the mined parallel data to get better vocabulary coverage and translation probability parameter estimation. Experiments are carried out on 10 years of Xinhua bilingual news collection. Using the mined data, we get significant improvement in word-to-word alignment accuracy in machine translation modeling.

1. Introduction

Parallel corpora are major resources for many natural language processing systems like cross-lingual information retrieval and statistical machine translation, which heavily rely on word-level translation equivalences. Existing parallel data resources available through organizations like the Linguistic Data Consortium are often limited in size, limited to major languages and not representative for specific tasks. Therefore, the Web is seen as an important source for bilingual corpora [2, 5, 6]. The focus so far has been to find parallel web pages, i.e. pairs of web pages where one is the translation of the other. For parallel corpora, which were created by translating existing text into some other language, a number of alignment approaches has been proposed [3, 5]. Using sentence length information or bilingual dictionaries proved efficient for this task.

Very often one can find data, which is not parallel in this strict sense but still closely related by conveying the same information. We call such data comparable. An example of such a resource is the news stories published by the Xinhua news agency in Chinese and English, which we used in our experiments. Building a sentence aligned bilingual corpus from such data is a much harder task than generating a sentence alignment for parallel data. The sentence alignment algorithm has not only to cope with sentences that have no translation at all, i.e. with insertions and deletions, but it also has to evaluate the sentence pairs how good translations they are. That is to say, we need to have an alignment model that gives reliable scores for all sentence pairs, which allows filtering out those that are not proper translations.

In our approach, a maximum likelihood criterion is proposed, which combines sentence length models and a statistical lexicon model. Specific models are formulated for insertions and deletions. The statistical lexicon can be extracted from an already existing sentence aligned parallel corpus. For sentence aligning the comparable corpus, which may be from a different domain, this poses the problem of low coverage resulting in less reliable alignment scores. To alleviate this problem we propose an iterative process that adapts the lexicon towards the new corpus, thereby giving higher vocabulary coverage and more reliable sentence alignment.

The paper is structured as follows: Section 2 presents the sentence alignment model. The translation lexicon based and sentence length based models are described in section 3 and 4. In section 5 experiments on a comparable bilingual news collection are demonstrated. Discussion and conclusions are given in section 6.

2. Alignment Model

Let S denote a news story in the source language (Chinese) and T a news story in the target language (English). Each story can be represented as a sequence of sentences as follows:

$$S = \{s_1, s_2, ..., s_j, ..., s_J\}, \qquad T = \{t_1, t_2, ..., t_i, ..., t_I\},$$

where s_j and t_i are sentence appearing in order in S and T respectively. The sentence alignment model is to calculate the distance between all possible parallel pairs (s_j, t_i) and to align them. The distance is based on both a translation lexicon and sentence length models. Dynamic Programming (DP) is applied to find the Viterbi path aligning the two sentence sequences in (S, T), and all aligned pairs are extracted and filtered from the alignment.

2.1. Maximum Likelihood Criterion

Let A denote the alignment between S and T. We want to find the sentence alignment $A^*_{[1:J][1:I]}$ that gives maximum likelihood of aligning S and T as follows:

$$A^*_{[1:J][1:I]} = \arg\max_A \{P(S:T \mid A)\} \qquad (1)$$

A consists of sub-alignments, $a_{(j,x)(i,y)} = \{[s_j, \cdots, s_{j+x}]:[t_i, \cdots, t_{i+y}]\}$, where x sentences in S are aligned to y sentences in T. Both x and y can be larger than 1 indicating one-to-many

0-7695-1754-4/02 $17.00 © 2002 IEEE

alignments, or zero indicating insertions/deletions. There are seven types allowed in our approach as defined in section 2.2. Under the assumption that the $a_{(j,x):(i,y)}$ are independent of each other, the probability in (1) can be approximated as follows:

$$
\begin{aligned}
P([1,s_J]:[1,t_I]\,|\,A) &= \prod_{a_{(j,x):(i,y)}\in A} P(a_{(j,x):(i,y)}\,|\,A) \\
&= \prod_{a_{(j,x):(i,y)}\in A} (P(A\,|\,a_{(j,x):(i,y)})P(a_{(j,x):(i,y)})/P(A)) \\
&= \prod_{a_{(j,x):(i,y)}\in A} (P(A\,|\,a_{(j,x):(i,y)})P([s_j\cdots s_{j+x}],[t_i\cdots t_{i+y}])/P(A)) \\
&= \prod_{a_{(j,x):(i,y)}\in A} P(A\,|\,a_{(j,x):(i,y)})P([s_j\cdots s_{j+x}]\,|\,[t_i\cdots t_{i+y}])P([t_i\cdots t_{i+y}])/P(A)
\end{aligned}
\qquad (2)
$$

We assume that all possible alignments A between story S and story T are equally probable, thus $P(A)$ is a const and can be omitted during the maximization in (1). Again, under the independence assumption regarding the $a_{(j,x):(i,y)}$, the maximization process can be implemented via a standard dynamic programming strategy depicted as follows:

$$
\begin{aligned}
A^*_{j+x,i+y} &= \arg\min_A\{-\log P([s_1\cdots s_{j+x}]:[t_1\cdots t_{i+y}]\,|\,A)\} \\
&\triangleq \arg\min_A\{D([s_1\cdots s_{j+x}]:[t_1\cdots t_{i+y}]\,|\,A)\} \\
&= \arg\min_A\{D([s_1\cdots s_{j-1}]:[t_1\cdots t_{i-1}]\,|\,A^*_{i-1,j-1}) + d([s_j\cdots s_{j+x}]:[t_i\cdots t_{i+y}]\,|\,A)\} \\
d([s_j\cdots s_{j+x}]:[t_i\cdots t_{i+y}]\,|\,A) &= -\log(P(a_{(j,x):(i,y)}\,|\,A)) \\
&= -\log(P(A\,|\,[s_j\cdots s_{j+x}])P([s_j\cdots s_{j+x}]\,|\,[t_i\cdots t_{i+y}])P([t_i\cdots t_{i+y}]))
\end{aligned}
$$

In (2), there are two types of probabilities: the *translation* probability $P([s_j:s_{j+x}]\,|\,[t_i:t_{i+y}])$ and *non-translation* probabilities: $P(A\,|\,a_{(j,x):(i,y)})$ and $P([t_i:t_{i+y}])$. These probabilities are to be approximated using lexicon-based models stated in section 3 and two sentence length models as stated in section 4.

2.2 Alignment types in Dynamic Programming

There are seven alignment types of $a_{(j,x):(i,y)}$ allowed in our dynamic programming approach as shown in Figure 1.

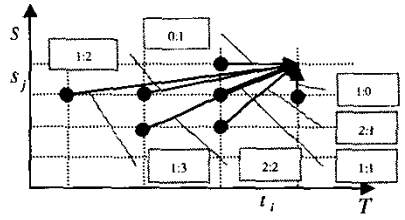

Figure-1 *Seven alignment types in DP*

The different alignment types (x:y) we use are: 1:1 (substitution), 1:0 (deletion), 0:1 (insertion), 1:2 (expansion), 2:1 (contraction), 1:3 (tri-expansion), and 2:2 (merge). For different language pairs and other corpora additional alignment types such as 3:1 and 3:3 might prove helpful.

3. Translation Lexicon Models

We used a statistical translation lexicon known as *Model-1* [1, 4] for both efficiency and simplicity. Detailed information can be found in [1].

Model-1 is the conditional probability that a word *sw* in the source language is translated given word *tw* in the target language: *t(sw|tw)*. The translation probability can be reliably estimated using the EM algorithm. The probability for an alignment of source sentence s given target sentence t is calculated as:

$$P(s\,|\,t) = \frac{1}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(sw_j\,|\,tw_i) \qquad (3)$$

where l and m are the corresponding sentence length measured as number of words in the sentences s and t.

In our approach, probability (3) is further normalized so that the probabilities for different lengths of sentence s are comparable at the word level:

$$\bar{P}(s\,|\,t) = \left[\frac{1}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(sw_j\,|\,tw_i)\right]^{1/m} \qquad (4)$$

This sentence alignment probability can be shown to reach global maximum [1], and leads itself to adaptation. We can use the mined parallel data to retrain *model-1* and update the alignment with better vocabulary coverage and better sentence alignment probability estimation in (4).

3.1. Explicit Alignments

The explicit alignments $a_{(j,x):(i,y)}$ are those alignments of which x and y are non-zero, including 1:2, 1:3, 2:2, 2:1, and 1:1. The calculation of the *translation* probability $P([s_j:s_{j+x}]\,|\,[t_i:t_{i+y}])$ in (2) is now explicit and straight forward. Let $s=[s_j:s_{j+x}]$ and $t=[t_i:t_{i+y}]$, and equation (4) is applied directly in this case.

3.2. Implicit Alignments: Background Models

Implicit alignments $a_{(j,x):(i,y)}$ are needed for deletions and insertions, where a text segment is aligned to Empty text. These correspond to the noises in the alignment, which have no translation counterpart in the comparable text. They are expressed by setting x or y zero.

One possibility is to align all words to an imaginary empty word "*NULL*", then apply (4). But this probability of aligning a word to *NULL* is very unreliable to be estimated; also both *t(fw|NULL)* and *t(NULL|ew)* need to be estimated, which will introduce more computation load and uncertainty; third, the length of the "empty text" in equation (4) is problematic to define.

In our approach, we build two background models, where the empty text is defined as all of the words in the vocabulary of a language. By this we actually assumed a large "sentence" consisting of all of the words in that vocabulary. The sentence length is then the vocabulary

size of that language. The insertion and deletion can now be defined as follows:

For deletion, the probability of aligning text s to empty text segment $\bar{P}(s \mid NULL)$ is defined via equation (4):

$$\bar{P}(s \mid NULL) = \left[\frac{1}{|V_T|^m} \prod_{j=1}^{m} \sum_{i=0}^{|V_S|} t(sw_j \mid tw_i) \right]^{1/m} \quad (5)$$

where $|V_T|$ is the vocabulary size of target language eg. English. For insertion, we have correspondingly:

$$\bar{P}(NULL \mid t) = \left[\frac{1}{(l+1)^{|V_S|}} \prod_{j=1}^{|V_S|} \sum_{i=0}^{l} t(sw_j \mid tw_i) \right]^{1/|V_S|} \quad (6)$$

Intuitively, common words in English are more likely to be insertions, and common words in Chinese are more likely to be deletions. Equations (5) and (6) are based on such intuitions.

By defining the empty text in this way, we can model insertion/deletion using the same equation as the other explicit alignment types. However, we expect that in (6), for a given tw_i, most of $t(sw_j \mid tw_i)$ can be very small. Therefore the product over the whole vocabulary V_S will be too small to be discriminative. In our approach, for a given tw_i, we only keep those $t(sw_j \mid tw_i)$ which are larger than a floor during our training of model-1.

4. Sentence Length Models

In equation (2), the two non-translation probabilities: $P(A \mid a_{(j,x)(i,y)})$ and $P([t_i : t_{i+y}])$ can be approximated by sentence length models, which were also tested on French-English [3]. Assuming the posterior probability of $P(A \mid a_{(j,x)(i,y)})$ is only related to the length of the aligned segments, we have:

$$\begin{aligned} P(A \mid a_{(i,x)(j,y)}) &= P(A \mid [s_j : s_{j+x}], [t_i : t_{i+y}]) \\ &\approx P(|[s_j : s_{j+x}]|:|[t_i : t_{i+y}]|, [s_j : s_{j+x}], [t_i : t_{i+y}]) \\ &\approx P(|[s_j : s_{j+x}]| - |[t_i : t_{i+y}]|) = P(\delta(x, y)) \end{aligned}$$

The length difference $\delta(x,y)$ is assumed to be a Gaussian distribution [3] and the normalized difference in equation (7) is considered to be normal distribution $N(0,1)$.

$$\bar{\delta} = \frac{y - x \cdot c}{\sqrt{(x+1)\sigma^2}} \quad (7)$$

where c is a const indicating the average length ratio between target and source sentences. For Chinese-English, x is dependent on word segmentation, and in our experiment c is 1.067. σ^2 is the variance of the sentence length difference, in our case, 0.197 shown in Table-1.

For the distribution of $P([t_i : t_{i+y}])$, one can use a target language model (n-gram) to approximate it. Here, to save the computation, we actually choose the Poisson distribution to model it. The Poisson distribution is a good approximation of the sentence length distribution and can be estimated from a monolingual corpus.

$$P([t_i \cdots t_{i+y}]) = \frac{\lambda_a^{|[t_i \cdots t_{i+y}]|}}{|[t_i \cdots t_{i+y}]|! e^{\lambda_a}} \quad (8)$$

λ_a is the expectation of the Poisson distribution. In our case, λ_a is the expectation of the length of segment $[t_i \cdots t_{i+y}]$. Following [5] we set λ_a to be different for different alignment types $a_{(j,x)(i,y)}$. $\lambda_{1:1}$ can be estimated using the English part of the Hong Kong news parallel corpus. The other λ_a are tied as follows:

$$\lambda_{1:0} = \lambda_{0:1} = \frac{1}{2}\lambda_{1:1} = \frac{1}{3}\lambda_{1:2} = \frac{1}{3}\lambda_{2:1} = \frac{1}{4}\lambda_{2:2} = \frac{1}{4}\lambda_{1:3}$$

Both (7) and (8) can be directly applied for insertion and deletion. In (7), x or y can be zero, indicating insertion and deletion. In (8), y can be zero, indicating deletion. Intuitively, a sentence pair with very large length difference is unlikely to be parallel, and abnormally short or long sentences are relatively more likely to be insertion/deletion (noise), as reflected in our sentence length models.

5. Experiments

Our work is motivated by the need to mine parallel sentences from the 10 years (1992~2001) Xinhua Web bilingual news corpora collected by Language Data Consortium (LDC). The collection is open-domain and comparable, with roughly similar sentence order of content. The English stories focus mainly on international news and the Chinese stories on domestic news. Most of the stories have no corresponding story in the other language. The first step was therefore to find comparable story pairs [2, 6]. This resulted in 17310 story pairs which were then used for the sentence alignment experiments. Each story has between 3 and 80 sentences, and the ratio of the number of sentences between English and Chinese stories is 1.36:1.

Preprocessing included word segmentation for the Chinese stories, separation of punctuation characters, and removal of web junk text. The Chinese full-stop '。' and the English period "." were used in sentence boundary detection. Table-1 shows the statistics of the sentence length models.

Table-1 *Sentence length model: word vs character*

	Word-based	Character-based
Mean	1.067	1.468
Var	0.197	0.275

The character-based model has a larger variance than the word-based model. No punctuations are counted in our sentence length models.

5.1. Parallel Sentence Alignment Models

First, we tested different alignment models, a character-based length model only (CL), a word-based length model only (WL), a translation model only (TB), and the

proposed maximum likelihood criterion combining WL and TB (WL/TB) as shown in Table 2.

Table-2: *Alignment types(%) in alignments resulting from different alignment models*

Models	0:1	1:0	1:1	1:2	2:1	2:2	1:3
CL	10.9	4.38	19.3	20.4	7.94	28.1	8.9
WL	4.63	2.99	57.5	18.3	3.45	5.11	8.05
TB	9.62	3.92	60.8	14.7	4.8	0.04	6.1
WL/TB	5.33	3.0	66.5	15.8	2.2	0.01	7.2

The length models (CL and WL) generate a large number of 2:2 alignments. While both TB and WL/TB give more reliable alignments from our close and detail examination. It showed necessary to incorporate the word translation identity information for robustness and accuracy. The combined WL/TB under maximum likelihood gives our best result.

5.2. Adaptive Parallel Data Mining

In a second experiment we used the mined data to re-train the translation lexicon. Results are shown in Table-3. There is no very large variation of the alignment type's proportion, but the aligned sentence-pair's perplexity changed significantly as shown in figure-2.

Table-3 *Alignment types (%) in adaptive extraction*

Iter	0:1	1:0	1:1	1:2	2:1	2:2	1:3
1	5.33	3.00	66.5	15.8	2.20	0.01	7.21
2	4.86	2.69	66.9	16.0	2.26	0.01	7.29
3	4.81	2.65	66.6	16.3	2.38	0.01	7.26
4	4.81	2.64	66.6	16.2	2.39	0.01	7.28

Table-4 *Vocabulary coverage of each iteration*

iteration	1	2	3	4
Chinese	65%	83%	85%	86%
English	57%	73%	80%	81%

There is a big increase of vocabulary coverage from iteration 1 to 2, shown in Table-4.

The sentence pair's perplexity *pp* is defined as:

$$pp = d([s_j : s_{j+x}] : [t_i : t_{i+y}] \mid A) \approx -\log(P(a_{(j,x)(i,y)} \mid A))$$

The distribution of the sentence-pairs perplexity in each iteration is shown in Figure 2. Sentence pairs, which have a perplexity less than the given threshold, are selected as training data for the next iteration.

Figure-2. *Distribution of aligned sentence pairs' perplexities*

Our approach got 110K (44MB) aligned parallel sentence-pairs, which are used to train word alignment models. The quality of the mined data is evaluated by word alignment accuracy of the models according to a gold-standard manually labeled test set.

5.3. Word-to-word Alignment Accuracy

A direct evaluation of the quality of mined parallel data is to test their effect in training word alignment models. We used GIZA++ [4] to build translation models up to model-3. Word alignment accuracy was then evaluated using a hand-aligned test set containing 365 sentence pairs with 4094 word-to-word alignments.

Table-5 *Word alignment of translation models*

Baseline	Model-1	Model-2	Model-3
Precision	43.43%	44.98%	43.65%
Recall	50.98%	53.81%	49.66%
F	46.90%	49.00%	46.46%
With mined data	Model-1	Model-2	Model-3
Precision	48.94%	48.88%	48.88%
Recall	58.97%	58.55%	56.84%
F	53.49%	53.28%	52.56%

The baseline models were trained using 290K sentence pairs from the Hong Kong News corpus available from LDC. The additional mined data was 57K sentence pairs (pp<5.0) selected after iteration 4. There is a consistent improvement for all three word alignment models. The harmonic mean F value of *Model-1* has a 14.05% relative improvement, showing better vocabulary coverage and high parallel quality of the data we mined.

6. Discussion and Conclusions

We described our approach of generating a high quality parallel corpus from a very large Chinese-English bilingual web text collection. Lexical information and sentence length information is combined to find reliable sentence alignment. The extracted parallel data was used to train word-based alignment models. The improved quality of the resulting word alignment proves the effectiveness of the sentence alignment method.

7. References

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

[2] Bing Zhao and Stephan Vogel, "Full-text Story alignment models for Chinese-English Bilingual Corpora", *International Conference on Spoken Language Processing*, Sep. 2002.

[3] Church, K. W. "Char_align: A Program for Aligning Parallel Texts at the Character Level". In *Proceedings of ACL-93*, Columbus OH. 1993.

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

[5] Stanley Chen. "Aligning sentences in Bilingual corpora using lexical information". In *Proceedings of the 31st Annual Conference of the Association for computational linguistics*, pages 9-16, Columbus, Ohio, June 1993.

[6] Xiaoyi Ma, Mark Y. Liberman, "BITS: A Method for Bilingual Text Search over the Web". *Machine Translation Summit VII*, 1999.