FILTERING NOISY PARALLEL CORPORA OF WEB PAGES

Jian-Yun Nie, Jian Cai

RALI lab., Département d'informatique et recherche opérationnelle
Université de Montréal, Montréal, Québec, H3C 3J7 Canada
{nie, cai}@iro.umontreal.ca

Abstract
In our previous study, we successfully built an automatic mining system for parallel texts from the Web – PTMiner that is able to determine a large number of parallel Web pages for different language pairs. However, there are a number of non-parallel text pairs in this corpus. This paper proposes a filtering approach to clean up the corpus. Our experiments show that once the corpus is cleaned, both the translation accuracy of the resulting translation models and the effectiveness of cross-language information retrieval (CLIR) using these models are improved significantly.

Keywords
Parallel text, statistical translation model, filtering, cross-language information retrieval

1. Introduction

As more and more non-English documents appear on the Web, there is a high demand for cross-language information retrieval (CLIR). Three approaches to CLIR are possible: using a machine translation (MT) system, using a bilingual dictionary or terminology database, or using a statistical translation model based on parallel texts.

Although MT systems seem to be a good tool for CLIR, we are restricted by its availability – no high-quality MT system exists for many language pairs. In addition, the translations by the current MT systems are not always reasonable.

Bilingual dictionary (or lexical databases) have been used in several CLIR experiments. In general, however, a simple use of such a resource leads to a low IR effectiveness.

The third approach based on parallel texts has some unique advantages:

- There is no need to acquire or to compile a bilingual dictionary or a complete MT system.
- Word translations are made sensitive to the domain, as embodied by the training corpus.

However, the problem with such an approach is often the unavailability of large parallel corpora for many language pairs. To solve this problem, in our previous research, we developed a system that automatically mines the Web for parallel texts – PTMiner. This system has successfully gathered a large number of "parallel" web pages for several language pairs – French-English, Chinese-English, Italian-English, and German-English. It has been shown [8] that they can be used to train statistical translation models for CLIR uses.

When we analyzed the constituted corpora more closely, we found that they still contain quite a number of non-parallel web pages. The goal of the current study is to clean the parallel corpus mined by PTMiner so that we can obtain better parallel texts and translation models. This paper reports the methods used in this filtering process, and the evaluation results in terms of both translation accuracy and CLIR effectiveness. Significant improvements have been observed with the filtered corpus in both evaluations. This work has been done for English-Chinese corpus. We believe that the same approach also apply to other language pairs.

In the following, we will first outline the principle of PTMiner. Then the filtering process will be described in detail. A series of evaluations will then be reported. Finally, we give some conclusions.

2. Principle of PTMiner

PTMiner [3] tries to determine parallel web pages through the following steps

1) Identification of candidate sites: The candidate web sites that may contain parallel pages are first identified. To do this, we take advantage of the existing search engines which have already indexed a large number of documents on the Web. For example, we send a query to AltaVista requesting for the English pages that contain anchor texts such as "Chinese version", "in Chinese", etc. Usually, such an anchor text is associated with a link to the Chinese version of the page. The sites filtered from the results are the candidate sites.

2) Search for web pages from the candidate sites: For each candidate site, we retrieve all the web pages that are indexed by the search engines (AltaVista and

Copyright © 2001

NorthernLight). From the set of documents found, we further browse the candidate sites thoroughly by following the links to other web pages on these sites.

3) Parallel page pairing: We use the following observation: parallel Web pages often have similar file names. For example, it is often the case that a file "file_e.html" in English often has a parallel page with the name "file_c.html" in Chinese. Only text pairs with similar names are kept.

4) Parallel page filtering: Some additional criteria are used to filter the pairs that are not parallel. For example, file length and language identification. If the lengths of two files are too different, they are not parallel. By using an automatic language identification [6], we can also filter out the pages that are not in the desired languages.

Finally, we obtained 14820 pairs of presumably parallel HTML pages.

3. Sentence alignment

Sentence alignment is crucial to the training of statistical translation model. Its objective is to find a mapping between sentences of the two texts.

Several alignment algorithms have been proposed in the literature, using length, cognates or lexical cues as criteria.

The algorithm of Gale & Church [5] is a typical statistical alignment algorithm based on sentence length. It determines the best alignments as follows:

Given paragraphs P_1 and P_2 in the parallel texts,

if length (P_1) ~ length (P_2) then

- *P1* and *P2* are considered to be possible match;
- For sentences S_i in P_1 and S_j in P_2, calculate their probability of match according to their length, in comparison with the typical length variance of the two languages,
- The final sentence alignments are those that give the maximal alignment probability to P_1 and P_2.

This method has been successfully used to align high quality parallel corpora. However, as pointed out by Simard et al. [9] and Chen [2], while aligning more noisy corpora, the methods based solely on sentence length are not robust enough. Simard et al. [9] proposed a method that uses lexical information - cognates - to help with alignment. Cognates are pairs of tokens of different languages, which share obvious phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. Examples are generation/génération and financed /financé for English/French.

The idea of cognate does not apply to Chinese/ English. Instead, Wu [10] incorporated a small lexicon of very frequent words as "lexical cues" into the length-based method. If lexical cues and their translations appear in the candidate sentences, this is an additional indication that the sentences are parallel.

4. Previous results and problems

Our previous work on the corpus constituted by PTMiner is as follows:

- Sentence alignment is based on a modified cognate-based algorithm of Simard et al. The modification considers that the HTML markups are special "cognates". This allowed us to consider the similarity in HTML structure between two texts.
- Once sentences are aligned, the 1-1 alignments are selected and used as the training data. This selection is done because most of the alignment errors occur in the cases of alignment of multiple sentences. The 1:1 alignments are the most reliable.
- Then two statistical translation models (C-E and E-C) are trained. These models are IBM model 1 [1].
- The translation models are used for query translation. For this, we submit each query to a translation model, and the N most probable translation words are selected as its translation (together with their probabilities). In our experiments, N is set to three times the length of the original query.

The main results are summarized as follows.

The training corpus produced by PTMiner system includes 117.2M Chinese text and 136.5M English text. Table 1 lists the statistics of the training source (after some pre-training process, e.g. HTML markup removal, Chinese word segmentation, etc.) for the testing model. There are 1,048,156 sentence alignments in total and 870,414 of them are 1:1 alignments.

	size	vocabulary	word count
English	74.1M	76,969	9,816,859
Chinese	51.1M	48,528	9,916,416

Table 1. Statistics of the training source

To test the translation accuracy, we randomly selected 200 words in Chinese and English, and examined the first translation of these words by the models. Table 2 gives the translation accuracy of the resulting translation models.

Copyright © 2001

E-C	C-E
81.5%	77%

Table 2. Translation accuracy of the models

The last evaluation is for cross-language information retrieval (CLIR), in which queries are translated with the translation models. We used two test corpora of TREC (http://trec.nist.gov), The Chinese corpus contains a set of newspaper articles from China. The English corpus contains articles of Associated Press. The Chinese corpus amounts to 170 Mb, and there are 54 queries. The size of the English corpus is 750Mb and there are 53 queries. Table 3 below gives a summary of the CLIR effectiveness in both cases. The effectiveness is measure by the 11-point average precision - a standard measure in information retrieval.

English monolingual	0.3861	Chinese monolingual	0.3976
C-E CLIR	0.1654 (42.8%)	E-C CLIR	0.1591 (40.0%)
C-E + dictionary	0.2583 (66.9%)	E-C + dictionary	0.2232 (56.1%)

Table 3. CLIR effectiveness using the models

The last row (C-E+dictionary and E-C+dictionary) corresponds to the combinations of the translation models with a small bilingual dictionary – CEDict [4]. The translations in the dictionary are added with a fixed value (0.005). We can see that when the query translation is complemented by a dictionary, the retrieval effectiveness is improved.

In comparison with CLIR effectiveness between European languages, which is typically around 80% of the monolingual effectiveness, the CLIR effectiveness between Chinese and English is relatively low. This is predictable because of larger difference between Chinese and English. The effectiveness shown in the above table is comparable to the results obtained by other researchers on the same data sets using a machine translation system [7]. This shows that statistical translation models are reasonable tools for query translation in CLIR.

In analyzing the training corpus, we observed that a certain number of texts are not truly parallel. Several particularities of web pages affected the results:

- There is no standard in naming web pages. Webmasters can name the files as they want. In some cases (in particular for files in Chinese), the names are not significant (e.g. numbers are often used as file names as in "010c.html"). Judging the parallelism of web pages according to name similarity will include errors.
- Even if PTMiner did some further checks, e.g. using file lengths and HTML tags, we allowed some degree of flexibility in order to not reject too many web pages. Therefore, a number of non-parallel texts still remain.

The question we raise in the current study is "how can we filter out some of the unparallel texts from the corpus so as to improve its quality?"

As the filtering process cannot be completely correct, it will eliminate some truly parallel text pairs along with non-parallel texts. Then another question is whether a smaller, but cleaner corpus results in a better translation models and better CLIR effectiveness than a larger but more noisy corpus. We will answer this question by conducting a series of tests.

5. Filtering principle

We used the following three criteria in the filtering process:

a) File length:

In the previous mining process, a loose filtering based on text length has been used. The difference in size is tolerated up to 40%. We observed that the length criterion has probably been applied with too much flexibility. Therefore, we try to enhance the length criterion, i.e. we require a text pair's length ratio to be closer to the typical length ratio between the two languages. Typically, the character length ratio between English and Chinese texts is about 2:1 [10]. If we consider that each Chinese character is stored in 2 bytes, then the size of a Chinese text is comparable to its English translation. The filtering on file length will only admit English-Chinese texts whose length ratio is within $[2-\delta, 2+\delta]$ where δ is the difference we tolerate. Notice that the previous filtering method used in PTMiner is equivalent to $\delta=0.8$.

b) Empty alignments

There was no further filtering at the sentence alignment step in the previous work. We observed that some unparallel texts could be easily eliminated if we examine the results of their sentence alignment – many sentences in them have not been successfully aligned, i.e. many 0:n and n:0 alignments. In our current filtering process, we eliminate those text pairs in which the proportion of such empty alignments is greater than a threshold.

c) Known translations

It is intuitive to us that parallel sentences should contain some "known" translations. In fact, for someone who is not familiar with both languages, but knows some words, the appearance of these known words and their translation in two sentences is a valid criterion to judge if the sentences can be parallel. In our process, we exploit this same principle. We use a

Copyright © 2001

bilingual dictionary as a set of "known" translations. When we observe that there is a known translation in the corresponding sentence, then the degree of correspondence between the two sentences is increased by a certain degree which is W*(proportion of known translations). This degree is integrated into the Gale&Church algorithm with the original probability based on length. So using the modified Gale&Church algorithm, we try to maximize the combined score. The importance W can be set at different value. We will test several values of W later. This principle is similar to "lexical cues".

The above three criteria can be combined together. In applying a single or combined criterion, if the criterion is not satisfied, then the text pair is judged unparallel and is eliminated from the corpus.

6. Evaluation of the filtering criteria

For our evaluations, we first set up a reference set of 484 pairs randomly selected from the original parallel corpus. We manually aligned this set of texts into parallel sentences, which correspond to the correct alignments. If we consider that texts with over 50% of sentences aligned are parallel, then 396 pairs (81.82%) of these texts are parallel. This number shows that there may be much improvement possible in this parallel corpus.

At the sentence level, the reference set contains 2 7923 (71.97%) non-empty alignments. Among them, 22 338 (57.63%) are 1-1 alignments.

Three metrics are used in our evaluations: precision, recall and F-measure. They are defined as follows:

$$precision = \frac{A}{B}$$

$$recall = \frac{A}{C}$$

$$F - measure = \frac{2 * precisoon * recall}{precisoon + recall}$$

where A = # correctly aligned items by the process,
B = # aligned items by the filtering process
C = # manually aligned items

When a strict filtering criterion is applied, usually precision increases, but recall drops. This is why we also use F-measure as a global measure, which is a kind of average between precision and recall.

6.1. Testing single criteria

The following table shows the filtering results in different settings of δ. We measure the three metrics for texts, non-empty sentence alignments and 1-1 sentence alignments. For texts, we consider again that a pair containing over 50% parallel sentences is parallel.

		δ			
		no filter	0.5	0.3	0.05
Text	P	0.66	0.72	0.79	0.86
	R	1.00	0.93	0.85	0.82
	F	0.79	0.81	0.82	0.84
Sent. align	P	0.67	0.73	0.80	0.87
	R	1.00	0.91	0.86	0.83
	F	0.81	0.81	0.83	0.83
1:1 align	P	0.72	0.79	0.86	0.91
	R	1.00	0.93	0.87	0.84
	F	0.84	0.86	0.86	0.88

Table 4. Filtering by length ratio

We can see that as the length criterion becomes stricter, the F-measure increases. This suggests that the global quality of the corpus is improved. However, the 0.05-difference with the typical length ratio seems to be a minimal value we have to tolerate, otherwise, the recall ratio will be too low.

The following table shows the effect of applying empty alignment criterion.

		Empty alignment <E			
Unit	measure	no filter	50%	20%	5%
Text	P	0.66	0.68	0.75	0.83
	R	1.00	0.98	0.86	0.82
	F	0.79	0.80	0.80	0.83
Sent. align	P	0.67	0.70	0.79	0.84
	R	1.00	0.98	0.87	0.83
	F	0.81	0.81	0.83	0.84
1:1 align	P	0.72	0.74	0.82	0.90
	R	1.00	0.96	0.87	0.84
	F	0.84	0.84	0.85	0.86

Table 5. Filtering by empty alignments

We also observe that when a stricter filtering is applied, better F-measure is obtained.

Now for the known translation factor, we used different importance value (W), and the results are described in Table 6.

		Weight for known translation (W)				
Unit	measure	no filt.	1.0	1.3	1.5	2.0
Text	P	0.66	0.69	0.76	0.85	0.70
	R	1.00	0.95	0.89	0.84	0.94
	F	0.79	0.80	0.82	0.84	0.80
Sent. align	P	0.67	0.70	0.77	0.86	0.70
	R	1.00	0.95	0.88	0.85	0.96
	F	0.81	0.81	0.82	0.86	0.81
1:1 align	P	0.72	0.76	0.81	0.90	0.76
	R	1.00	0.95	0.91	0.87	0.94
	F	0.84	0.84	0.85	0.88	0.84

Table 6. Filtering by known translations

Copyright © 2001

We see that we cannot attribute a too large weight to the known translation factor. The best value (among those tested) seems to be 1.5. This suggests that there should be a balance between the length factor in the original Gale&Church algorithm, and the known translation factor.

In order to see the impact of the size of the dictionary, we conducted the following experiment: we select subsets of N most frequent words and their translations from the dictionary, and measure the alignment accuracy. The alignment accuracy is defined as follows:

$$alignment_accuracy = \frac{\#correct_alignments}{\#alignments}$$

The following table shows the results.

Dictionary size	alignment accuracy
no dictionary	72.31%
500 words	72.95%
1 000 words	74.36%
3 000 words	77.89%
5 000 words	80.82%
8 000 words	81.78%
10 000 words	82.16%

Table 7. Impact of the size of the dictionary

We can see that a small dictionary (500 words) does not have much impact on the alignment. With a large dictionary, the impact is significant. Once the size of the dictionary is reasonably large (5 000 and more), the improvements due to the increases in dictionary size become smaller.

In contrast with [10] which only used a few hundreds common words and their translations as lexical cues, our experiments suggest that using a larger dictionary is better.

6.2. Combining different criteria

There are many possible ways to combine the three criteria. In our test, we combined them as follows:

First, the length criterion is applied to filter texts. Only text pairs with [2-δ, 2+δ] range are kept.

Then the modified Simard algorithm – which uses HTML tags as cognates, and integrates the known translation factor, is applied to find the best alignments.

Finally, we examine the alignment results, and eliminate text pairs in which the proportion of empty alignments is higher than a threshold.

The following table shows two very different cases – one tries to have a higher recall and another a higher precision.

		Combinations (E+δ+W)	
		50%+0.5+1.0	5%+0.05+1.5
Text	P	0.74	0.92
	R	0.95	0.81
	F	0.83	0.87
Sent. align	P	0.75	0.93
	R	0.95	0.82
	F	0.84	0.87
1:1 align	P	0.80	0.97
	R	0.94	0.83
	F	0.86	0.87

Table 8. Combination of the filtering criteria

We see that a stricter filtering (targeting at a higher precision) leads to a better F-measure. This seems to suggest that a higher precision is preferable to a higher recall for our corpus. However, this remains to be proven. So our next test tries to see if this is true by evaluating the translation accuracy and the CLIR effectiveness.

6.3. Translation quality

To test the translation quality of a translation model, we used the same set of 200 randomly selected words. The translation accuracy is defined as follows:

$$translation_accuracy = \frac{\#correct_translations}{200}$$

The following table gives the translation accuracies of the two cases.

Direction	Combination		
	No filter	0%+0.5+1.0	5%+0.05+1.5
E-C	161 (80.50%)	172 (86.00%)	183 (91.50%)
C-E	154 (77.00%)	163 (81.50%)	173 (86.50%)

Table 10. Translation accuracy

We can see that the translation accuracy is consistent with F-measure, i.e. it is preferable to have a higher precision than a higher recall. We believe that this is due to the reasonably large size of our original corpus. So, even a number of text pairs have been eliminated, the remaining pairs are still enough for model training.

In comparison with the translation accuracy before filtering, the (5%+0.05+1.5) case represents an improvement of respectively 12.27% and 14.94% for E-C and C-E.

6.4. CLIR experiments

Our last series of tests is on CLIR. We use the translation models, together with the CEDict dictionary, for query translation, as described in

Copyright © 2001

Section 4. The following two tables describe the C-E (Chinese queries and English documents) and E-C CLIR.

Method	Trans. accuracy	Avg_prec. (%mono)	With dict. (%mono)
50%+0.5+1.0	86%	0.1898 (49.16%)	0.26.89 (69.64%)
50%+0.05+1.0	88.5%	0.1952 (50.56%)	0.2736 (70.86%)
5%+0.05+1.0	90%	0.2009 (52.03%)	0.2765 (71.61%)
5%+0.05%+1.5	91.5%	0.2063 (53.43%)	0.2811 (72.81%)

Table 9. Chinese-English CLIR effectiveness

Combined method	Trans. accuracy	Avg_prec. (%mono)	With dict. (%mono)
50%+0.5+1.0	81.5%	0.1843 (47.11%)	0.2463 (61.95%)
50%+0.05+1.0	83.5%	0.1921 (48.31%)	0.2498 (62.83%)
5%+0.05+1.0	85.5%	0.1976 (49.70%)	0.2537 (63.81%)
5%+0.05%+1.5	86.5%	0.2013 (50.63%)	0.2601 (65.42%)

Table 11. English-Chinese CLIR effectiveness

Observing the changes along with different combinations, we see that the CLIR effectiveness is also consistent with F-measure and translation accuracy: a better F-measure and translation accuracy result in a better CLIR effectiveness. This shows that a stricter filtering of the training corpus is also beneficial for CLIR.

The combination with the dictionary is also consistent with the previous result: this combination always leads to a better effectiveness.

If we compare these results with those obtained with the unfiltered corpus (Table 3), we can see that significant improvements have been achieved: 24.72% and 26.51% improvements for E-C and C-E CLIR without dictionary, and 8.83% and 16.53% for E-C and C-E CLIR with dictionary.

7. Conclusions

In this paper, we proposed an approach to filter a noisy parallel corpus. Our goal is to eliminate non-parallel texts from the corpus. In so doing, we hope to improve both the translation accuracy of the resulting translation models, and the CLIR effectiveness using these models.

Our filtering approach is based on the following three factors: length, known translations and empty alignments. Our experiments clearly showed that all these three factors contributed in improving the quality of the corpus.

The experiments on translation accuracy and CLIR effectiveness showed that it is preferable to have a strict filtering in our case. This may be due to the large size of our original corpus. So even a certain number of truly parallel texts are filterer out, there are still enough texts for model training. A strict filtering also led to a better CLIR effectiveness.

The filtering process can be further improved. For example, there are other possible combinations of the filtering criteria, e.g. combining them directly within the alignment algorithm. There may also be additional criteria. So in the future, we still exploit more filtering approaches to further improve the quality of the corpus. We will also apply the filtering process to similar corpora of other language pairs – French-English, Italian-English and German-English.

References

1 P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of machine translation: Parameter estimation. Computational Linguistics, 19: 263-311, 1993.

2 S. F. Chen. Aligning sentences in bilingual corpora using lexical information. Proc. ACL, pp. 9-16, 1993.

3 J. Chen, J.Y. Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proc. ANLP, pp. 21-28, Seattle, 2000.

4 P. Denisowski. Cedict project, http://www.mindspring.com/~paul_denisowski/cedict.html, 1999.

5 W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. Proc. ACL, pp. 177-184., 1991.

6 P. Isabelle, G. Foster, and P. Plamondon. SILC: a System for Language and coding identification. http://www-rali.iro.umontreal.ca/ProjetSILC.en.html, 1997.

7 K. L. Kwok and L. Grunfeld, TREC-5 English and Chinese retrieval experiments using PIRCS. Proc. TREC-5, NIST SP 500-238. Ed. Harman, D. K. and Voorhees, E. M, pp. 133-142, 1996.

8 J.Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining parallel texts from the Web. Proc. ACM SIGIR, pages 74-81, 1999.

9 M. Simard, G. F. Foster, and P. Isabelle. Using cognates to align sentences in bilingual corpora. Proc. TMI-92, 1992.

10 D. Wu. Aligning a parallel English-Chinese corpus statistically with lexical criteria. Proc. ACL, pp. 80-87, 1994.

Copyright © 2001