Champollion: A Robust Parallel Text Sentence Aligner

Xiaoyi Ma

Linguistic Data Consortium
3600 Market St. Suite 810
Philadelphia, PA 19104
xma@ldc.upenn.edu

Abstract

This paper describes Champollion, a lexicon-based sentence aligner designed for robust alignment of potential noisy parallel text. Champollion increases the robustness of the alignment by assigning greater weights to less frequent translated words. Experiments on a manually aligned Chinese – English parallel corpus show that Champollion achieves high precision and recall on noisy data. Champollion can be easily ported to new language pairs. It's freely available to the public.

1. Introduction

Parallel text is a very valuable resource for a number of natural language processing tasks, including machine translation (Brown et al. 1993; Vogel and Tribble 2002; Yamada and Knight 2001;), cross language information retrieval, and word disambiguation.

Parallel text provides the maximum utility when it is sentence aligned. The sentence alignment process maps sentences in the source text to their translation. The labor intensive and time consuming nature of manual sentence alignment makes large parallel text corpus development difficult. Thus a number of automatic sentence alignment approaches have been proposed and utilized; some are pure length based approaches, some are lexicon based, and some are a mixture of the two approaches.

While existing approaches perform reasonably well on close language pairs, such as English and French, their performance degrades quickly on remote language pairs such as English and Chinese. Performance degradation is exacerbated by noise in the data; we will explore the effects of noise in the data in the section 3.

Our research towards a robust sentence aligner on remote language pairs and noisy data produced Champollion, a lexicon-based sentence aligner. Champollion was initially developed for aligning Chinese-English parallel text. It was later ported to other language pairs, including Arabic – English and Hindi – English.

The rest of the paper is laid out as follows: Section 2 gives a brief overview of previous work; Section 3 describes the challenge of sentence aligning large parallel text corpora; Section 4 describes the algorithm of Champollion; Section 5 describes the experiments and the results; and Section 6 concludes the paper.

2. Previous Work

The first attempt to automatically align parallel text was (Gale and Church 1991), which is based on the idea that long sentences will be translated into long sentences and short sentences into short ones. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences and the variance of this difference. The probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences.

The length based approach works remarkably well on language pairs with high length correlation, such as French and English. Its performance degrades quickly, however, when the length correlation breaks down, such as in the case of Chinese and English.

Even with language pairs with high length correlation, the Gale-Church algorithm may fail at regions that contain many sentences with similar length. A number of algorithms, such as (Wu 1994), try to overcome the weaknesses of length based approaches by utilizing lexical information from translation lexicons, and/or through the identification of cognates.

In addition to the length based approach and length and lexicon hybrid approach, there are a few other approaches in the literature.

(Chen 1996) builds a sentence-based translation model and find the alignment with the highest probability given the model.

(Melamed 1999) first finds token correspondences between the source text and its translation by using a pattern recognition method. These token correspondences are used in conjunction with segment boundary information to find sentence correspondences.

3. Challenge from Noisy Data

Most data mentioned in the literature are relatively clean. For example, 1-1 alignment constitutes 89% of the UBS English-French corpus in (Gale and Church 1991), and 1-0 and 0-1 alignments constitute merely 1.3%.

However, when creating very large parallel corpora, the data can be very noisy. In particular, we see more 1-0 and 0-1 alignments than we previously did. For example, in a UN Chinese English corpus, 6.4% percent of all alignments are either 1-0 or 0-1 alignment (See Table 1).

Category	Frequency	Percentage
1-1	1306	89.4%
1-0 or 0-1	93	6.4%
1-2 or 2-1	60	4.1%
others	2	0.1%
total	1461	

Table 1: types of alignment in a sample UN corpus

Some of the omissions and insertions were introduced during the translation of the text. Most of the omissions and insertions, however, are introduced during different stages of processing before sentence alignment is carried out. In the UN case, we collected 12 years worth of articles which totaled about 60,000 documents in each of the 7 UN official languages. The raw documents were in three different formats (unknown format, WordPerfect, and MS Word), each had 4 to 5 minor versions. The layouts of the documents in different languages were different, and the layout of a specific language also changes over time. The pre-processing steps include converting the raw data to plain text format, removing tables, foot notes, end notes, etc. Most of these steps introduce noise. For instance, while a table in an English document can be completely removed, this is not necessarily the case in any given Chinese document.

Because of the sheer number of documents involved, manually examining each document after pre-processing is impossible. A robust sentence aligner needs not only to detect most categories of noise, but also to recover quickly if an error is made.

Our study shows that existing methods work very well on clean data, but their performance goes down quickly as data becomes noisy.

4. Champollion

Champollion differs from other sentence aligners in two ways. First, it assumes a noisy input, i.e. that a large percentage of alignments will not be one to one alignments, and that the number of deletions and insertions will be significant. The assumption is against declaring a match in the absence of lexical evidence. Non-lexical measures, such as sentence length information – which are often unreliable when dealing with noisy data – can and should still be used, but they should only play a supporting role when lexical evidence is present.

Second, Champollion differs from other lexicon-based approaches in assigning weights to translated words.

Translation lexicons usually help sentence aligners in the following way: first, translated words are identified by using entries from a translation lexicon; second, statistics of translated words are then used to identify sentence correspondences.

In most existing sentence alignment algorithms, translated words are treated equally, i.e. translated word pairs are assigned equal weight when deciding sentence correspondences.

Should these translated word pairs have an equal say about whether two sentences are translations of each other? Probably not. For example, assume that we have the following sentence pair in a report on recent waves of violence in Iraq:

a. Marketplace bombing kills 23 in Iraq
b. 伊拉克 集市 爆炸 造成 23 人 死亡

The translation pair (23, 23) is much stronger evidence than (Iraq, 伊拉克) that the two sentences are a match, simply because "Iraq" and "伊拉克" appear much more often than "23".

Assigning greater weight to less frequent translation pairs is the basis of Champollion. Champollion uses a function to compute the similarity between any two segments, each of which consists of one or more sentences. There is a penalty associated with alignments other than 1-1 alignment. The penalty is determined empirically. Sentences with a mismatching length are also penalized.

Champollion then uses a dynamic programming method to find the optimal alignment which maximizes the similarity of the source text and the translation.

4.1. Similarity Function

Champollion borrowed the idea of *tf-idf* weight (term frequency - inverse document frequency), which is often used in Information Retrieval, to compute the similarity (or relevancy) of two segments (one segment contains one or more sentences), one in the source language, the other in the target language. The reasoning was that the fundamental problem of sentence alignment is to score the relevancy of one segment in the source text and one in the translation. If we treat segments as documents, we reduce the problem of scoring segment relevancy to scoring document relevancy in IR.

We define *stf* as the segment-wide term frequency, i.e. the number of occurrences of a term within a segment, and *idtf* as the inverse document-wide term frequency, which can be computed by

$$idtf = \frac{T}{\#occurences\_in\_the\_document}$$

where *T* is the total number of terms in the document.

The *stf* gives a measure of the importance of the term within the particular segment. The *idtf* is a measure of the general importance of the term in the document.

The *stf–idtf* measure evaluates the importance of a translated word pair is to the alignment of two segments. The importance increases proportionally to the number of times a word appears in the segment but is offset by how common the word is in the entire document.

Champollion treats segments as bags of words, word ordering is not considered.

Assume that we have two segments, *E* and *C*, defined as follows:

$$E = \{e_1, e_2, \ldots, e_{m-1}, e_m\}$$
$$C = \{c_1, c_2, \ldots, c_{n-1}, c_n\}$$

where e_i and c_j are word tokens.

We define the *k* translated word pairs identified between the two segments as

$$P = \{(e'_1, c'_1), (e'_2, c'_2) \ldots (e'_k, c'_k)\}$$

then similarity of *E* and *C* is defined as

$$sim(E,C) = \sum_{i=1}^{k} \lg(stf(e'_i, c'_i) * idtf(e'_i)$$
$$* alignment\_penalty_{ij}$$
$$* length\_penalty(E,C)$$

where *alignment_penalty* is 1 for 1-1 alignment and a number between 0 and 1 for other kinds of alignments; *length_penalty* is a function of the length of the source segment and the length of the target segment.

4.2. The Dynamic Programming Algorithm

The dynamic programming algorithm is very similar to (Gale and Church 1991). However, instead of searching for the path with the minimum distance, we search for the path with maximum similarity. Champollion allows 1-0, 0-1, 1-1, 2-1, 1-2, 1-3, 3-1, 1-4 and 4-1 alignment.

We compute a lattice *S(i, j)* representing the similarity from the beginning of the document to the *i*th source sentence and *j*th target sentence. This lattice can be calculated efficiently using a simple recurrence relation:

$$S(i,j) = \max \begin{cases} S(i-1,j) + sim(Seg_{i,i}, \phi) \\ S(i,j-1) + sim(\phi, Seg_{j,j}) \\ S(i-1,j-1) + sim(Seg_{i,i}, Seg_{j,j}) \\ S(i-1,j-2) + sim(Seg_{i,i}, Seg_{j-1,j}) \\ S(i-2,j-1) + sim(Seg_{i-1,i}, Seg_{j,j}) \\ S(i-2,j-2) + sim(Seg_{i-1,i}, Seg_{j-1,j}) \\ S(i-1,j-3) + sim(Seg_{i,i}, Seg_{j-2,j}) \\ S(i-3,j-1) + sim(Seg_{i-2,i}, Seg_{j,j}) \\ S(i-1,j-4) + sim(Seg_{i,i}, Seg_{j-3,j}) \\ S(i-4,j-1) + sim(Seg_{i-3,i}, Seg_{j,j}) \end{cases}$$

where $Seg_{a,b}$ represents all segments numbered between *a* and *b*, inclusively.

In other words, the most probable alignment of the first *i* source sentences and *j* target sentences can be expressed in terms of the most probable alignment of some prefix of these sentences extended by a single alignment. The rows in the equation correspond to 1-0, 0-1, 1-1, 1-2, 2-1, 1-3, 3-1, 1-4 and 4-1 alignment, respectively.

4.3. Tokenizers

Champollion tokenizes both sides of the parallel text before computing the optimal alignment. For morphologically complicated languages, such as English and Arabic, this involves first splitting sentences into words by white space, and then applying a light stemmer. The stemmer is used to normalize the words to their dictionary forms, and thus maximizes the number of lexical matches.

Chinese text doesn't have explicit word boundaries. A word segmenter is used to tokenize Chinese sentences. We compared the performance of Champollion with and without word segmentation and found word segmentation improves both precision and recall.

5. Evaluation

We evaluated the performance of Champollion on a manually aligned Chinese – English parallel text corpus.

We also ran experiments to study the impact of translation lexicon coverage on Champollion performance.

We did not to run other sentence aligners to compare their performance with Champollion; however, the evaluation corpus and the gold alignment that we produced are part of the Champollion package, which is open source and freely available to the public. Those who are interested can download Champollion and run their aligners on the same set of data to compare their performance with Champollion.

5.1. Evaluation Corpus

The evaluation corpus was selected from three Chinese – English parallel corpora: Sinorama Magazine (Taiwan), Hong Kong Hansard, and the United Nations official documents. Table 2 shows the makeup of the evaluation corpus.

	Sinorama	HK Hansard	UN
# documents	2	3	2
# english sentences	462	1,927	1,399
# chinese sentences	412	1,961	1,493

Table 2: Makeup of evaluation corpus

The evaluation corpus was sentence segmented, and then manually aligned. The initial alignment was performed by a native Chinese speaker who is also fluent in English. The result was then double checked by another fluent bilingual.

There were a few cases where a Chinese sentence was only partially translated into English. The annotator was instructed to mark a pair as a match only if the majority of the words were translated. These cases were rare, and it would not have had a significant impact on the evaluation if they were treated differently.

Table 3 shows the statistics of the gold alignment by source and by alignment type. Overall, 1-1 alignment accounts for only 81.6% of the evaluation corpus, while deletions and insertions (1-0 and 0-1 alignments) constitute 6.3% of the corpus. These numbers are typical of a noisy parallel text corpus.

		Sinorama	HK Hansard	UN	**Total**
1-1	alignment	228	1,473	1,306	3,007
	percentage	56.3%	81.1%	89.4%	81.6%
1-0 0 1	alignment	74	65	93	232
	percentage	18.3%	3.6%	6.4%	6.3%
1-2 2 1	alignment	79	248	60	387
	percentage	19.5%	13.6%	4.1%	10.5%
2-2	alignment	2	8	0	10
	percentage	0.5%	0.4%	0.0%	0.3%
1-3 3 1	alignment	12	20	1	33
	percentage	3.0%	1.1%	0.1%	0.9%
1-4 4 1	alignment	9	1	1	11
	percentage	2.2%	0.1%	0.1%	0.3%
other	alignment	1	2	0	3
	percentage	0.2%	0.1%	0.0%	0.1%
Total		405	1,817	1,461	3,683

Table 3: Statistics of the gold alignment

5.2. Experiments

The English – Chinese translation lexicon used in the experiment contains about 58,000 head words. The lexicon was created by combining a few bilingual English – Chinese dictionaries collected from the Internet. Very limited quality control was done on the translation lexicon, but in general, the translation lexicon was clean.

To study the impact of dictionary coverage on the performance of Champollion, we created artificial translation lexicons of different sizes by 1) generating a sorted frequency list of English words computed over a large English corpus; 2) selecting the most frequent K words from the top of the list; 3) extracting dictionary entries associated with these K words. K is the number of head words in the resulting dictionary. In this experiment, 1,000, 2,000, 4,000, 8,000, 16,000, 28,000 and 58,000 were the values for K.

5.3. Results

Table 4 shows the precisions and recalls for different runs where translation lexicons of different sizes were used. Using a lexicon of 58K head words, Champollion achieved 97.0% precision and 96.9% recall on the test set.

#head words	Precision	Recall
0	0.881	0.908
1K	0.936	0.934
2K	0.954	0.953
4K	0.964	0.963
8K	0.963	0.963
16K	0.965	0.964
28K	0.965	0.964
58K	0.970	0.969

Table 4: Precision and recall by dictionary size

Note that champollion achieved 88.1% precision and 90.8% recall without using a translation lexicon. This is because champollion also uses tokens which remain the same when being translated, these include punctuations, numbers, abbreviations, such as “IBM”.

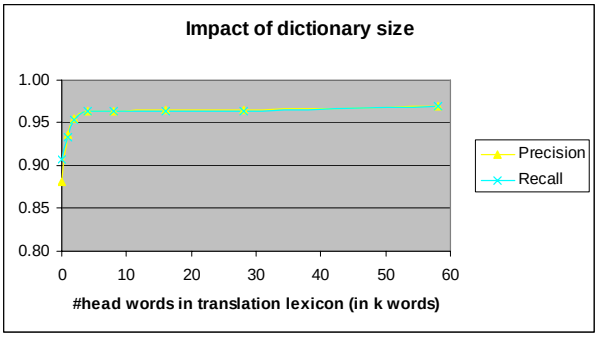

Figure 1: Precision and recall by dictionary size

Figure 1 is a plot of the same numbers presented in Table 4. X coordinate is the size of the translation lexicon, Y coordinate is the precision and recall. The curves show that the performance of Champollion reaches a plateau using a lexicon with about 10K head words. It also indicates that the performance gain from using very large translation lexicons seems small.

Table 5 shows the precisions and recalls by alignment type, aligned by using the full size (58K head words) lexicon. The numbers indicate that Champollion gets almost all the 1-1 alignment right. Its precision and recall for other types of alignment are not as good.

Type	Precision	Recall
1-1	0.977	0.970
1-0 and 0-1	0.546	0.453
1-2 and 2-1	0.578	0.584
2-2	0.353	0.600
1-3 and 3-1	0.643	0.818
1-4 and 4-1	0.733	1.000
others	0.000	0.000

Table 5: Precision and recall by type

6. Conclusion

This paper describes Champollion, a lexicon-based parallel text sentence aligner. Champollion considers a match possible only when lexical matches are present. It assigns higher weight to less frequent words, which are considered a stronger indication that two segments are a match. Sentence length information is used to weed out bogus matches.

Champollion achieved high precision and recall on manually aligned Chinese-English parallel text corpus. Champollion and the evaluation data used in this paper are available at http://champollion.sourceforge.net.

7. References

Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. (1993). The mathematics of statistical Machine Translation: Parameter estimation. *Computational Linguistics* 19, 2, pp. 263-311.

Chen, S. (1996). Building Probabilistic Models for Natural Language. *Ph.D. dissertation*, Harvard University, Cambridge, MA.

Gale, W. A., Church, K. W. (1991). A program for aligning sentences in bilingual corpora', in *ACL '91*, Berkeley CA, pp. 177-184.

Melamed, I.D. 1999. Bitext maps and Alignment via Pattern Recognition. *Computational Linguistic*, 25, No 1, pp 107-130.

Wu, D.. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the*

32nd Annual Meeting, pages 80-87, Las Cruces, NM. Association for Computational Linguistics.

Vogel, S. and Tribble, A. (2002). Improving statistical machine translation for a speech-tospeech translation task. In *Proceedings of the 7th International Conferenc on Spoken Language Processing (ICSLP-02).*

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Anniversary Meeting of the Association for Computational Linguistics (ACL-01).*