Full-text Story Alignment Models for Chinese-English Bilingual News Corpora

Bing Zhao, Stephan Vogel

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

ABSTRACT

In this paper, we describe the full-text story alignment on Chinese-English bilingual corpora of news data to mine potential parallel data for machine translation. Several standard information retrieval methods are tested and two translation-model based alignment models are proposed and studied. Modeling the process of generating the parallel English story from Chinese story gives significant improvements over the standard information retrieval techniques. Refinements of the alignment model are also proposed and tested in detail. On one day's bilingual news collection, our methods improved the mean reciprocal rank from 0.31 to 0.68.

1. INTRODUCTION

Parallel data are a major resource for many NLP systems such as machine translation, cross lingual information retrieval, and lexical acquisition.

The XinHua news agency, the largest news agency in Mainland China, publishes its news stories on the WWW every day in Chinese and English. There are several hundred stories published every day in both languages. But Chinese and English stories are not parallel in the sense of being translations of each other. First, most of Chinese stories focused on domestic news, while English news stories focused on the international news. Second, Chinese and English usually describe the same story in very different ways and styles.

Finding those story pairs that are translations or at least near translations is difficult for a number of reasons. The vocabulary of the XinHua data is pretty large, including many proper names such as person's name etc. Therefore, available dictionaries and translation systems give very low vocabulary coverage. In addition, the language pair of Chinese and English is very different in terms of grammar and linguistic properties. These characteristics combined together pose a great challenge for mining the parallel data from this bilingual collection of XinHua news stories.

A few of approaches aiming at cross-lingual information retrieval (IR) have been proposed. An example is in [3], which combined manually and probabilistic lexicons by a weighting scheme. In [2] it has been shown that using a probabilistic translation model gives results comparable to the direct translation of the query using on-line public machine translation systems. In [6], a similarity metric based on co-occurrence counts of translation pairs within a local text window is proposed to calculate the distance between parallel texts. For linguistically similar language pairs such as German and English, these approaches showed good performance.

These methods have proved the effectiveness of using a probabilistic translation model in cross-lingual retrieval. But most of these IR methods are dealing with query in one language and retrieving documents in another language, and the methods are applied to the linguistically similar language pairs such as French, German and English.

In our approach, we first investigate the standard information retrieval techniques for parallel stories alignment by creating pseudo-queries from Chinese stories. Then two translation lexicon based alignment models are proposed. The refinements and improvements of the alignment models are also discussed.

This paper is structured as follows: in section 2, translation model is briefly introduced. In section 3, pseudo-query generation and the standard information retrieval techniques are introduced; in section 4, two translation-model based alignment models are described. The refined alignment model is described in section 5. The experiment, comparison and discussion are given in section 6 and 7.

2. TRANSLATION MODEL

In this paper, we used a statistical machine translation model similar to [1], and Model-1 [1] is used for its simplicity and efficiency.

2.1. Translation Model

In brief, Model-1 is the conditional probability that a word f in source language is translated given word e in target language, t(f|e). Given training data, which consists of parallel sentences: $\{(f^{(i)}, e^{(i)}), i = 1..S\}$, Model-1 computes:

$$t(f \mid e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f \mid e; f^{(s)}, e^{(s)}) \tag{1}$$

where λ_e^{-1} is a normalization factor such that $\sum_f t(f \mid e) = 1$

$$c(f \mid e; f^{(s)}, e^{(s)}) = \frac{t(f \mid e)}{\sum_{k=1}^{l} t(f \mid e_k)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i) \tag{2}$$

where $l = | f^s |, m = | e^s |$. Thus an EM algorithm is applied to train Model-1. For detailed information, see [1].

2.2. System description

The training data used in our system is Hong Kong news parallel corpora provided by Language Data Consortium (LDC). There are 290K parallel sentences. During training, we isolated all the punctuations and English words are all lowercased. The Chinese word segmentation are based on the LDC Chinese-English dictionary 2.0, thus each segmented

Chinese word has English translations in the dictionary. The vocabulary statistics we obtained from HongKong news corpus and from XinHua collection are shown in Table-1:

Table-1: vocabulary coverage of the XinHua Corpora

Vocab	XinHua	HongKong	Overlap	Percent
Chinese	45807	151959	24563	54%
English	66129	134658	45542	69%

The vocabulary coverage of the XinHua collection is quite low. For English, it is 69% coverage, and for Chinese, only 54%. The vocabulary mismatch is a common problem that all systems have to face when mining data from the web. In our translation model, we assign each unseen word a default translation probability.

We trained two probabilistic translation models: one is from English into Chinese, and another is from Chinese into English. The two models are used for the two alignment models described in section 4 and 5.

3. PSEUDO QUSERY MODEL

A pseudo-query is essentially a bag of English words translated from Chinese words with translation probabilities over a given threshold or simply top N-best translation words. Each Chinese story is then a query, and the potential parallel English news story is the relevant document to be retrieved.

3.1. Generation of the pseudo queries

In our approach, for each Chinese word c_i in a Chinese story: $\{c_i, i = 1..m\}$, we take its top 3 translation words for c_i: (e_i^1, e_i^2, e_i^3) from our translation model, then take union of all the translated English words $\{(e_i^1, e_i^2, e_i^3), i = 1..m\}$, remove stop words, and a pseudo-query is generated.

Now the story alignment reduces to the evaluation of the pseudo-query and the retrieval of one particular document, which is the potential parallel English story (none or only one parallel story). This is a typical known-item retrieval task stated in [5], where we know there is only one correct answer for a query. We are interested in retrieving that one particular document instead of retrieving/ranking the entire set of documents that pertain to a particular subject like in standard IR.

3.2. Retrieval methods

We here used three standard query methods. They are TFIDF, Okapi, and LM-JM [4]. The first two methods differ in term weighting schemes. Okapi has the advantage of the term frequency normalization using document length:

TFIDF: $tw = [\log(tf) + 1)]$

Okapi: $tw = tf / [tf + 0.5 + 1.5 \cdot doclen / avg(docLen)]$

The query *q* and document *d* are represented by term-weight vectors, of which the element is an index term's weight *tw* in *q* and *d*. The similarity of *q* and *d* is the inner product of the two vectors.

The third method of LM-JM is a language modeling approach to information retrieval. It infers a language model for each document and estimates the generation of the query according to each of the document language models. Then it ranks the documents using those probabilities. In this paper, the document language model (LM) is built as a unigram and the smoothing method is Jelinek-Mercer (JM), which is fixed coefficient linear interpolation.

4. ALIGNMENT MODEL

4.1. Alignment Model-A

Instead of using pseudo-queries, we can directly model the probability that an English story (Document) D_E was the relevant document given a Chinese story (Query) Q_C : $P(D_E | Q_C)$. Assume story is a bag of tokens, we then have:

$$D_E = \{w_e^i, i = 1..l\}, \; Q_C = \{w_c^i, i = 1..m\}$$

And $P(D_E | Q_C)$ is formalized via a translation model t(e|c) as follows:

$$\begin{aligned} P(D_E | Q_C) &= \prod_{w_e \in D_E} \left[\sum_{w_c \in Q_C} P(w_e, w_c | Q_C) \right] \\ &= \prod_{w_e \in D_E} \left[\sum_{w_c \in Q_C} P(w_e | w_c, Q_C) \cdot P(w_c | Q_C) \right] \\ &\doteq \prod_{w_e \in D_E} \left[\sum_{w_c \in Q_C} P(w_e | w_c) \cdot P(w_c | Q_C) \right] \end{aligned} \tag{3}$$

where $P(w_e | w_c) = t(w_e | w_c)$ is the translation model, which is trained as stated in Section 2. $P(w_c | Q_C)$ is in fact a language model, which gives the probability that the word w_c is generated from query Q_C. In our approach, it is a unigram language model.

Thus, the relevance of an English story to the given Chinese story (which is a query) is directly modeled. For each Chinese news story, we can obtain a ranked candidates list with a score of relevance attached to each of the candidates. The translation probability acts as a bridge between the language pairs in an efficient way. It re-weighs each possible English word's relevance to the given Chinese story. The lexicon lookup of the probability $t(w_e | w_c)$ is very quick.

In our approach, we also expanded 10K top frequent w_c in a separate memory block, and this speed up the process by a factor of 3.

4.2. Alignment Model-B: A Generative Model

Another way of doing the parallel story alignment is to model the probability that a Chinese Story (Query) Q_C is *generated* from an English Story (Document) D_E: $P(Q_C | D_E)$, which is a generative model. Assume each story is a bag of tokens. Similar to Alignment Model-A, we have:

$$\begin{aligned} P(Q_C | D_E) &= \prod_{w_c \in Q_C} \left[\sum_{w_e \in D_E} P(w_c, w_e | D_E) \right] \\ &= \prod_{w_c \in Q_C} \left[\sum_{w_e \in D_E} P(w_c | w_e, D_E) \cdot P(w_e | D_E) \right] \\ &\doteq \prod_{w_c \in Q_C} \left[\sum_{w_e \in D_E} P(w_c | w_e) \cdot P(w_e | D_E) \right] \end{aligned} \tag{4}$$

where the translation model $P(w_c | w_e) = t(w_c | w_e)$ is trained using English as the source language and Chinese as the target language. $P(w_e \mid D_E)$ is the document language model of generating the word w_e from D_E, a unigram language model:

$$P(w_e \mid D_E) = count\,(w_e, D_E)\,/\,length\,(D_E)$$

This approach is different from alignment Model-A, in that this alignment model assumes a generation process, in which an English Story generates a Chinese story with a probability.

5. REFINED ALIGNMENT MODEL

5.1. A sliding window

Both models have one common problem: ignoring the position of the words appearing in the documents (stories). And all the stories are deemed as a bag of words. We know that potential parallel stories have a good portion of parallel sentences, and token pairs in these sentences usually appear in similar positions of the corresponding stories. For example, if a token w_c is too far away from the position of token w_e, it is unlikely for them to be a translation token pair.

While it is hard to model the position of translation pairs, we can approximate the effect by using a window to constraint the probabilities of translation for the token pair. The refinement uses alignment Model-B for demonstrating this formalization, which is actually the same for Model-A. The refinement of alignment Model-B is an approximation as follows:

Assume stories are tokens with positions:

$$D_E = \{(w_e, i), i = 1..l\},\ Q_C = \{(w_c, i), i = 1..m\}$$

$$\begin{aligned}
P(Q_C \mid D_E) &= \prod_{(w_c,i)\in Q_C}\left[\sum_{(w_e,j)\in D_E} P((w_c,i),(w_e,j) \mid D_E)\right] \qquad (5)\\
&\doteq \prod_{w_c\in Q_C}\left[\sum_{w_e\in D_E} P((w_c,i) \mid (w_e,j))\cdot P((w_e,j) \mid D_E)\right]\\
&\cong \prod_{w_c\in Q_C}\left[\sum_{w_e\in D_E} P(w_c \mid w_e)P(i \mid j)\cdot P(w_e \mid D_E)\right]\\
&\cong \prod_{w_c\in Q_C}\left[\sum_{w_e\in D_E} P(w_c \mid w_e)P(|i-j|)\cdot P(w_e \mid D_E)\right]
\end{aligned}$$

where $P(|i-j|)$ is the probability of the position difference $|i-j|$ between the translation token pair (w_c, w_e). But this probability is hard to model in our approach; we actually use a uniform probability within a certain size of window:

$$P(|i-j|) = \begin{cases} 1/(2*N), if\ |i-j| < N \\ 0, \qquad else \end{cases} \qquad (6)$$

where N is the window size.

5.2. The TFIDF approximation

Another refinement is the $P(w_e \mid D_E)$, which is a unigram document language model.

$$P(w_e \mid D_E) = \frac{c(w_e, D_E)}{|D_E|} = \frac{tf_{w_e}}{|D_E|} \qquad (7)$$

where tf_{w_e} is the term frequency in document D_E and $|D_E|$ is the length of D_E. Actually, we can increase the discrimination power by replacing $|D_E|$ with the document frequency within the whole collection of English stories. Thus $P(w_e \mid D_E)$ is approximated by TFIDF as follows:

$$P(Q_C \mid D_E) \cong \prod_{w_c\in Q_C}\left[\sum_{w_e\in D_E} P(w_c \mid w_e)P(|i-j|)\cdot \frac{(1+\log(tf_{w_e}))(\log N/df_{w_e})}{\lambda}\right] \qquad (8)$$

where N is the number of English documents in the collection, and λ is a normalization const. The TFIDF captures the words' discrimination power from the whole collection's statistics by using term inverse document frequency (IDF), thus is using more information than the document language model.

In our approach, we further normalized the probability as follows:

$$\mathrm{PP} \hat{=} -\frac{1}{|\mathrm{Q_C}|}\log P(Q_C \mid D_E) \qquad (9)$$

Thus good parallel story pairs will have a low normalized PP. This normalized score can help us to set thresholds in later experiments.

6. EXPERIMETAL RESULTS

For manual evaluation we selected the stories from one day in the year of 2000 XinHua bilingual news collection. There are 214 Chinese stories and 168 English stories. We hand-labeled the data, and found 26 pairs of parallel stories of good quality. The parallel story alignment can be considered as known-item retrieval [5]: each Chinese story has either none or exactly one parallel English story. The known-item retrieval approach is well-suited to this task, as we stated in section 3. Two evaluation measures can be applied for this kind of problem:

- Rank of the known-item in an N-best list;
- Mean-reciprocal rank measure, which is the mean of the reciprocal of the rank at which the known-item was found.

The first one gives very close and detail evaluation, and the second one gives a single value summary, which is in fact the average precision. We use these two measures together in our experiments.

Standard pseudo-query retrieval is carried out using the Lemur Toolkit [4]. The pseudo-query is generated using the top-3 translation words from each word in a Chinese story.

Two translation models are trained using the Hong Kong news data, one from English to Chinese and vice versa for alignment Model-A and Model-B, respectively. The simple TFIDF Okapi and LM-JM are compared with the alignment models. The evaluation is shown in Figure 1.

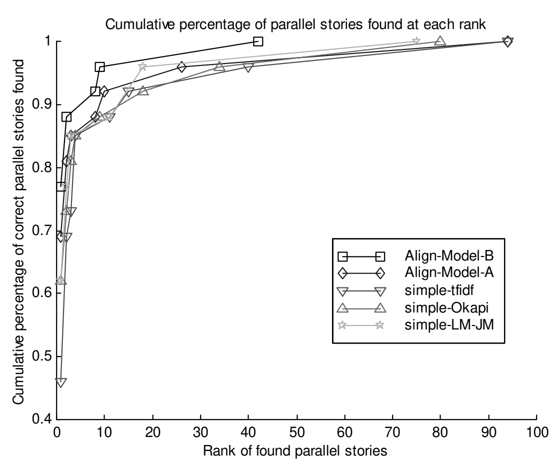

Figure 1 *Cummulative percentage of parallel stories found up to given rank*

Table 2 gives the numbers of correctly retrieved items at ranks up to 3. The mean-reciprocal rank is also shown in the bottom row.

Table 2 Mean reciprocal rank and known items at raw rank.

Rank	M-A	M-B	TFIDF	Okapi	LM-JM
1	18	20	12	16	14
2	21	23	18	19	18
3	22	23	19	21	19
$1/\bar{r}$	0.158	0.306	0.131	0.150	0.171

From Figure 1 we see that alignment Model-B achieves the best performance. Alignment Model-A also achieves good performance when looking at the top 3 ranks. This confirms that the translation-model based alignment models, which directly model the relevance between the Chinese stories and the English stories, are more reliable than the pseudo-query approach. The pseudo-query is actually a noise channel, and the information of the Chinese stories is lost when passing through this channel, hurting alignment performance. For alignment models, Model-B is better than Model-A. This is because Model-A is a product of all the words' probability in the English story, and this product is very sensitive to the length of the story as shown in Equation (3). Long English story will in general have a small probability due to the many products of probabilities. For the pseudo-query approach, the simple LM-JM is better than Okapi and TFIDF, and it is also slightly better than Model-A in terms of the mean-reciprocal rank. This is because LM-JM implicitly performs document length normalization, while Model-A and simple TFIDF do not. But in general, the pseudo-query models have similar performance in terms of mean-reciprocal rank.

A second experiment was carried out to investigate the different refinements for alignment Model-B. First, the effect of the window size as given in Equation (6) was tested. We checked the top-1 result against the normalized scores shown given in Equation (9). The result is shown in Table 3.

Table 3 Top-1 results for different window sizes and PP thresholds.

PP\N	20	30	40	50	60
PP<5	0	0	2/2	2/3	3/5
PP<6	3/3	4/6	9/16	18/35	18/72
PP<7	10/22	19/122	17/195	19/213	19/214
$1/\bar{r}$	0.206	0.302	0.321	0.342	0.406

In each cell of Table 3, the nominator is the number of correctly retrieved known-items, and the denominator is the total number of retrieved documents within the window size and PP range, two thresholds to obtain better performance. The window size encoded the translation word pair's position information. From observation in our experiment, a window size of 20-word corresponds approximately to a sentence, and 50-word corresponds to a paragraph. With too long a window size, tokens will be too far apart from each other to be translation pairs, thus the performance of the alignment model is hurt. Our experiment showed that with a window size of 40~50 words, and a normalized PP of less than 5.0, most of the potential parallel stories come out within the top 3 ranks.

The second improvement in alignment Model-B is the removal of all punctuations and English stop words. These words are not informative in finding parallel stories, and actually they are aligned to too many Chinese words, which is not desirable. The removal of the punctuations and stop words can reduce confusion across English stories, and also reduce the computation load for these words. The experiment result is shown in Table 4.

Table 4 Top-1 result: Removal of the punctuations & stopwords

PP\N	20	30	40	50	60
PP<5	0	2/2	4/6	4/6	4/8
PP<6	3/3	5/10	13/29	19/66	21/111
PP<7	16/39	22/142	21/206	22/214	22/214
$1/\bar{r}$	0.329	0.394	0.542	0.578	0.542

The third improvement is to use a TFIDF weight instead of the unigram language model probability. The result is shown in Table-5.

Table 5 Top-1 result: using TFIDF weight

PP\N	20	30	40	50	60
PP<5	1/1	3/3	5/8	6/12	6/15
PP<6	3/3	6/13	22/98	21/132	21/170
PP<7	18/40	22/148	22/214	22/214	22/214
$1/\bar{r}$	0.329	0.520	0.650	0.684	0.578

By using TFIDF, further discrimination is achieved. Term inverse document frequency is calculated from the whole collection, using more information than document language models. With these three improvements, we achieved our best performance at a window size of 50 words. The mean reciprocal rank was 0.684.

7. CONCLUSION AND FUTURE WORK

In this paper, we showed how statistical alignment models can be used to find parallel stories in a large bilingual collection. Additional refinements and improvements of the alignment models were discussed and tested on the Chinese-English bilingual news collection. The experiments showed that modeling the generation process of the parallel English story from the Chinese story with a statistical alignment model significantly outperforms standard information retrieval approaches. Future work will be done in modeling the news stories' titles and combining them with the full-text alignment model to obtain even more reliable and better parallel story alignment performance.

8. 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", *Computation Linguistics* 19(2): 263-311 1993.

[2] Jian-Yun, Nie, Michel Simard, Pierre Isabelle, Richard Durand, "Cross Language Information Retrieval based on Parallel Text and Automatic Mining of Parallel text from Web", *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[3] Jinxi Xu, Alexander Fraser, Raph Weischedel. "TREC 2001 Cross-lingual Retrieval at BBN". *The tenth Text Retrieval Conference*. 2001.

[4] Lemur Tookit: See http://www-2.cs.cmu.edu/~lemur/.

[5] Paul B. Kantor, Ellen Voorhees, "Report on the TREC-5 Confusion Track". *The Fifth Text Retrieval Conference*, 1996.

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