Filtering or Adapting: Two Strategies to Exploit Noisy Parallel Corpora for Cross-Language Information Retrieval

Lixin Shi Jian-Yun Nie
Dept d'Informatique et de Recherche Opérationnelle, Université de Montréal
C.P. 6128, succursale Centre-ville, Montréal, Québec, H3C 3J7 Canada
shilixin@iro.umontreal.ca nie@iro.umontreal.ca

ABSTRACT
Noisy parallel corpora have been widely used for Cross-language information retrieval (CLIR). However, the previous studies only focus on truly parallel corpus. In this paper, we examine two possible approaches to exploit noisy corpora: filtering out noise from the corpora or adapting the training process of translation model to the noise corpora. Our experiments show that the second approach is better suited to CLIR.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Query Formulation

General Terms
Algorithms, Theory, Experimentation, Performance

Keywords
Parallel Corpus, Noise Filtering, Alignment Score.

1. INTRODUCTION
Parallel corpora have been used as a means of query translation in cross-language information retrieval (CLIR). The previous methods consist of first training a translation model (TM) and then applying it to query translation. They usually assume that the corpora are truly parallel, even if they are automatically mined from the Web [6, 7]. In reality, the corpora are often noisy: part of them is not parallel. Then an acute question is how to take into account the noise during the exploitation of the corpora.

Some previous studies have tried to filter out noise from a parallel corpus [3, 5] before model training. However, the conclusions vary: [5] found that noise filtering is useful to improve both translation quality and CLIR effectiveness, while [3] did not find it useful for machine translation. In this paper, we will re-examine noise filtering from the CLIR perspective. In addition, we will also consider an alternative approach, which takes into account the noise (via alignment score) during the training of translation model. Our experiments will show that noise filtering can improve the resulting CLIR effectiveness if the filtering threshold is set correctly. On the other hand, by integrating alignment score into model training process, we can obtain results comparable to that of filtering but without running the danger of over-filtering.

2. RELATED WORK
The exploitation of a parallel corpus usually follows the following steps: sentence alignment, translation model training from the aligned sentences [1]. However, a common assumption is that the given training corpus is truly parallel, which is not the case for those automatically mined from the Web [6, 7]. Therefore, noise filtering becomes a necessary step. [5] uses file length, empty alignments and known translation (according to a bilingual dictionary) to filter out noise text pairs, while [3] considers sentence length and translation likelihood as filtering criteria. The above two studies seem to draw contradictory conclusions: [5] reports positive impact, while [3] reports almost no impact with noise filtering. The difference between them may be due to their different goals: one focuses on CLIR and another on machine translation. In this paper, we will re-examine the exploitation of noisy corpora from the CLIR perspective.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011..

3. FILTERING VS. ADAPTING
A noisy corpus can contain two types of noise: completely unparallel texts (i.e. two unrelated texts) and partially unparallel texts (i.e. part of them is parallel). Here, we mainly deal with the second case, because in our text mining step, we can apply quite strict criteria to filter out the first type of noise [6].

Sentence filtering using Sentence Alignment Score

To filter out noise, [3] used sentence alignment score, which is determined according to sentence length and translation likelihood. As translation likelihood is also determined from the training corpus alone, the alignment score may not be a reliable measure of the noisiness of a sentence pair. To improve the accuracy of the estimation, we use a bilingual dictionary: if two sentences contain many mutual translation words, there is a high chance that the sentences are parallel. According to this criterion, we determine the following score:

$$Score(\mathbf{e},\mathbf{f}) = \frac{\#translation\ words\ between\ \mathbf{e}\ and\ \mathbf{f}\ by\ a\ dictionary}{(\#\ words\ in\ \mathbf{e} + \#\ words\ in\ \mathbf{f})/2}$$

This score is then integrated into the alignment process as follows: Given a sequence S of source sentences and a sequence T of target sentences, we try to determine the alignment A such that:

$$\arg\max_A \Pr(A \mid S,T) \approx \arg\max_A \prod_{k=1}^{K} P(B_k^{(A)})$$

$$\approx \arg\max_A \sum_{k=1}^{K} \log[P(\alpha\_align)P(\delta \mid \alpha\_align)Score(\mathbf{e},\mathbf{f})]$$

where $B_1,...,B_K$ are sentence beads under alignment A; α_align is a match of type α (such as 1:0, 1:1, 2:1); $P(\delta \mid \alpha\_align)$ is determined according to the length ratio between two alignment candidates, as in [2].

Training TM by Considering Sentence Alignment Scores

In IBM models, one assumes that the training corpus is clean and only contains correctly aligned sentences. Each sentence alignment is given the same importance during the training process of translation model. Now, if we know that some of the sentence pairs are noise, we should rely less on them for the

training. Here, we assume that the alignment score $p^{(s)}$ ($0 \le p^{(s)} \le 1$) reflects the parallelism of the sentence pair. Then we can modify the EM algorithm of IBM model 1 [1] by integrating the alignment score as follows:

E-Step: Compute the count of alignment between (e, f)

$$c'(e \mid f) = \sum_{s=1}^{S} c'(e \mid f; \mathbf{e}^{(s)}, \mathbf{f}^{(s)}, p^{(s)})$$

$$= \sum_{s=1}^{S} [p^{(s)} \frac{t(e \mid f)}{\sum_{j=0}^{m} t(e \mid f_j)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i)]$$

M-Step: Compute the translation probability $t(f|e)$

$$t(e \mid f) = c'(e \mid f) / \sum_{e} c'(e \mid f) .$$

It is expected that the resulting model relies more on strongly aligned (presumably true) sentence pairs, and less on weakly aligned (noisy) sentence pairs.

4. EXPERIMENTS

In order to contrast the impact of the above approaches on noisy and non-noisy corpora, we use the following two parallel corpora: one is a non-noisy corpus from the Hong Kong Government Information Centre[1] (Corpus A), another one is noisy, which is mined from six websites from the United Nations, Hong Kong, Taiwan and Mainland China (Corpus B). After sentence alignment, we obtain 700,000 pairs of parallel sentences in Corpus A, and 281,000 pairs in Corpus B.

The test collection of CLIR is as follows: Documents are Chinese documents used in TREC6 and queries are English queries CH1-CH54. We use "title" and "description" of topics. For Chinese document indexing, we use both words and characters.

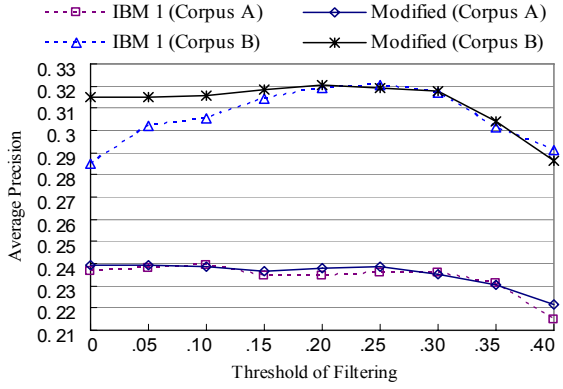

Figure 1.The results of CLIR of using different TMs

We first filter out noise according to alignment score (keeping those that are higher than a threshold). Then TMs are trained using the traditional IBM model 1 or our modified model. For query translation, we keep 6 strongest translation words in Chinese for each English query word. Figure 1 shows the CLIR results using different TMs.

From the figure, we first observe that the effectiveness of CLIR depends on the coverage of the training corpus. Corpus A contains more sentence pairs and has higher translation quality. However, it can only suggest translation words in a narrow area. Corpus B from six websites is of smaller size, but its coverage is larger. As a consequence, the CLIR effectiveness using Corpus B is better than Corpus A. This observation shows that the coverage of the training corpus is crucial.

As expected, noise filtering does not have noticeable impact on the non-noisy corpus (A), except that it can reduce the effectiveness by over-filtering (when threshold>0.25).

The most important observation in this study is that our modified model performs usually better than the traditional IBM model 1, especially when the filtering threshold is low. With the modified model, even without filtering, we can obtain CLIR effectiveness comparable to the best one with filtering. This shows that the modified training process can naturally account for the noise in the training corpus, and filtering becomes then optional. In addition, another advantage of the modified model is that with no filtering or with a light filtering, we do not run the risk of removing too many true parallel pairs (over-filtering) and reducing the CLIR effectiveness (when the threshold>0.30). This suggests that the modified model is a better solution to account for noise in parallel corpora than noise filtering.

5. CONCLUSION

In this paper, we investigated two possible ways to exploit noisy corpora: either filtering out as much noise as possible, or modifying the model training process so as to take into account the noise in the corpora. Our results show that both approaches can lead to some improvements in CLIR effectiveness based on noisy corpora. However, when we incorporate the sentence alignment score (or noisiness) into model training, we can arrive at similarly good CLIR effectiveness without running the danger of over-filtering. So, this approach seems to be a better one for the exploitation of noisy corpora for CLIR.

6. REFERENCES

[1] Brown, P.F., S.A. Della Pietra, V. J. Della Pietra, and R.L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comp. Ling.* 19(2):263-311, 1993.

[2] Gale, W. A. and K.W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Comp. Ling.* 19(1): 75-102, 1993.

[3] Khadivi, S. and H. Ney. Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. *NLDB,* pp. 263-274, 2005.

[4] Kraaij, W., J.Y. Nie, and M. Simard. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. *Comp. Ling.* 29(3): 381-419, 2003

[5] Nie, J.Y. and J. Cai. Filtering noisy Parallel Corpora of Web Pages. *IEEE symp. on NLPKE*, pp.453-458, 2001.

[6] Nie, J.Y., M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. *SIGIR,* pp.74-81, 1999.

[7] Resnik, P. Mining the Web for Bilingual Text. *ACL*, pp.527-534, 1999.

[1] http://www.info.gov.hk