A Language Modeling Approach for Extracting Translation Knowledge from Comparable Corpora

Razieh Rahimi and Azadeh Shakery

School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
{razrahimi,shakery}@ut.ac.ir

Abstract. A main challenge in *Cross-Language* information retrieval is to estimate a translation language model, as its quality directly affects the retrieval performance. The translation language model is built using translation resources such as bilingual dictionaries, parallel corpora, or comparable corpora. In general, high quality resources may not be available for scarce-resource languages. For these languages, efficient exploitation of commonly available resources such as *comparable corpora* is considered more crucial. In this paper, we focus on using only comparable corpora to extract translation information more efficiently. We propose a *language modeling* approach for estimating the translation language model. The proposed method is based on probability distribution estimation, and can be tuned easier in comparison with heuristically adjusted previous work. Experiment results show a significant improvement in the translation quality and CLIR performance compared to the previous approaches.

Keywords: Cross-language Information Retrieval, Translation Language Models, Comparable Corpora.

1 Introduction

Cross-Language Information Retrieval refers to retrieval process where documents and queries are in different languages. Crossing the language barrier is an additional vital step of CLIR in comparison with monolingual information retrieval. Consequently, we need some kind of translation which can impose a limitation on the CLIR performance. The goal in CLIR is to eliminate the language barrier and make the CLIR performance comparable to monolingual IR performance, but crafting this purpose highly depends on translation qualities and appropriate usage of the translation knowledge. This turns the attention to the translation resources with this rule of thumb in mind: wider coverage and higher quality of translation resources lead to superior translations.

Various resources have been used for obtaining translation knowledge for CLIR, including machine translation systems, bilingual dictionaries and thesauri, and bilingual parallel or comparable corpora. A *Parallel corpus* is a collection

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 606–617, 2013.

of document pairs that are exact translations of each other. Parallel corpora have been widely used for translation extraction with effective CLIR results for different language pairs [13]. However, obtaining such high quality resources is time-consuming and costly. Lack of parallel corpora can result in low-quality machine translation systems. Therefore, these translation resources may not be available for resource-lean languages. In contrast, a *Comparable corpus* is a document collection in which aligned documents cover the same or similar topics. Comparable corpora can be built with less cost because of its looser alignments. Extracting translation knowledge from comparable corpora is more challenging than parallel corpora due to noisy alignments. But, in case of resource limitation, we need to efficiently utilize comparable corpora in extracting translations.

Several approaches for extracting translation knowledge from comparable corpora are proposed [3, 4, 6, 7, 9–12, 15–21]. Most of these approaches employ additional lexical resource(s) besides comparable corpora. Guiding the process of translation extraction from comparable corpora or combining the resulting translations by translations from other resources has positive impact on translation qualities and improves the CLIR performance. However, focusing on using only comparable corpora is valuable by itself. Methods that improve translation knowledge extraction only from comparable corpora are regarded as crucial to performance of cross-language retrieval concerning minority languages. In addition, better translations from comparable corpora can be combined with other translation resources for resource-rich languages. In this paper, we focus on extracting translation knowledge from only comparable corpora.

Exploiting only comparable corpora for translation extraction has been considered in a few studies [17–20]. These approaches represent each word by a vector, and some of them use heuristics in calculation of vector elements [18,19]. Numerous factors can influence the quality of extracted translations, such as: source/target document length normalization, the ratio of source to target document length, and length normalization of source/target word vectors. In addition, these factors are dependent on the source/target language characteristics as well as the attributes of the comparable corpus. Using heuristics to address these factors increases the number of parameters that require tuning. Experiments show that using different heuristics affects the CLIR performance significantly. But, investigating all cases to find the optimal settings can be impractical.

In this paper, we propose a more principled method for extracting translations from comparable corpora. We adopt the *Language Modeling* approach for monolingual information retrieval to translation extraction problem. The intuition behind our proposed method is that words that are translations of each other have similar contributions in generating language models of the aligned documents. This method improves the quality of extracted translations and related words in the target language. Indeed, the proposed approach can be optimized straightforwardly compared to methods that use heuristics.

The paper is organized in four parts. In Section 2, we briefly describe the approaches used for comparison. Then, we present our proposed language modeling method for translation extraction in Section 3. Following experimental

design and the results are reported in Section 4. Finally, the paper is concluded in Section 5.

2 Translation Extraction for CLIR

In this section, two previous methods for obtaining translation language models from only comparable corpora are discussed. A comparable corpus consisting of n alignments which are represented as *source document id* (d_{s_i}), *target document id* (d_{t_i}), and *alignment similarity* (s_i) triples, is formulated as:

$$A = \left\{(d_{s_1}, d_{t_1}, s_1), .., (d_{s_n}, d_{t_n}, s_n)\right\} . \tag{1}$$

Since alignments from source to target documents of comparable corpora have many-to-many correspondence, the documents d_{s_i} and d_{s_j} (similarly d_{t_i} and d_{t_j}) in which $i \neq j$ may be the same. $d_{s_i} = d_{s_j}$ means that the original document is duplicated when it is aligned with different target language documents.

To build the translation language model, similarity scores between source-target word pairs are needed. Let w_s and w_t represent a word in the source and target language respectively. The goal is to calculate the similarity score between w_s and w_t in order to find the most similar words in the target language to w_s.

2.1 Frequency Correlation-Based Approach

The approach proposed in [20] extracts translations based on correlation between frequency distributions of terms. The more correlated the term distributions in the comparable corpus, the higher the similarity scores. Formally, source and target word vectors are $w_s = \{x_1, ..., x_n\}$ and $w_t = \{y_1, ..., y_n\}$ respectively, where $x_i = \frac{\mathbf{tf}(w_s, d_{s_i})}{\sum_{j=1}^{n} \mathbf{tf}(w_s, d_{s_j})}$, $y_i = \frac{\mathbf{tf}(w_t, d_{t_i})}{\sum_{j=1}^{n} \mathbf{tf}(w_t, d_{t_j})}$ and n is the length of the alignment vector in Eq. 1. The similarity of w_s and w_t is measured using Pearson's correlation coefficient as:

$$\mathbf{sim}(w_s, w_t) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{N} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)}} . \tag{2}$$

2.2 Cocot Approach

This approach, proposed in [19], exploits *tf.idf* weighting scheme for extracting word translations by reversing the role of documents and words. In this approach, target documents that are aligned with a same source document are grouped into a hyper document. So, the word vectors are built according to $A = \{(d_{s_1}, D_{T_1}), ..., (d_{s_m}, D_{T_m})\}$ alignment vector where D_{T_i} is a hyper document. Source and target term vectors are $w_s = \{x_1, ..., x_m\}$ and $w_t = \{Y_1, ..., Y_m\}$ respectively, where:

$$x_i = \begin{cases} 0 & \text{if } w_s \in d_{s_i} \\ (0.5 + 0.5\frac{tf(w_s, d_{s_i})}{\max tf(d_{s_i})}).\ln(\frac{NT}{|d_{s_i}|}) & \text{otherwise} \end{cases} , Y_i = \sum_{d_{t_j} \in D_{T_i}} \frac{y_j}{\ln(\text{rank}_{d_{t_j}} + 1)} , \tag{3}$$

and NT is the number of unique terms in the collection. In Cocot approach, cosine normalization is used for source word vectors, while pivoted length normalization is adopted for target word vectors. Finally similarity of two words is calculated as the inner product of corresponding vectors:

$$\mathbf{sim}(w_s, w_t) = \frac{\sum_{\langle d_{s_k}, D_{T_k}\rangle \in A} x_k Y_k}{\|\boldsymbol{w}_s\|\big((1-\text{slope}) + \text{slope}\frac{\|\boldsymbol{w}_t\|}{\text{avg-trg-vlength}}\big)} \ . \tag{4}$$

2.3 Using Translation Language Model in CLIR

In his survey paper [13], Nie presents two approaches for integrating word translations in language modeling approach using the KL-divergence retrieval model. In our work, we use "Query Translation approach" which builds a new query model that incorporates the translation probabilities. In this approach for CLIR, documents are ranked based on:

$$\begin{aligned} \text{score}(Q, D) &= \sum_{w_t \in V_t} P(w_t|\theta_Q) \log P(w_t|\theta_D) \ , \\ p(w_t|\theta_Q) &= \sum_{w_s \in V_s} P(w_t|w_s) P(w_s|\theta_Q) \ , \end{aligned} \tag{5}$$

where θ_Q and θ_D are query and document language models respectively, and $P(w_t|w_s)$ indicates the probability that the source word w_s is translated to the target word w_t. To apply this approach, these probabilities should be estimated using the similarity scores of translation extraction approaches.

Suppose we choose top m translations $w_{t_1}, ..., w_{t_m}$ for a source word w_s from extracted translations, with similarity scores $s_1, ..., s_m$ respectively. We use *Naive normalization* approach to estimate $p(w_{t_i}|w_s)$, according to which we have:

$$p(w_{t_i}|w_s) = \frac{\mathbf{sim}(w_s, w_{t_i})}{\sum_{j=1}^{m} \mathbf{sim}(w_s, w_{t_j})} \ . \tag{6}$$

3 Language Modeling Approach for Translation Extraction

In *Language Modeling* approach, a document language model represents the word distribution from which the document is sampled. The basic idea of exploiting this approach for translation extraction is that words that are translations of each other have similar contributions in word distributions of aligned documents. In the first step, we represent each word by a model that captures the contribution of the word in the language model of the document in each alignment. Then, the similarity is measured based on the KL-Divergence between source-target word models. Formally the translation extraction process is as:

$$D(\theta_{w_s} \| \theta_{w_t}) = \sum_{i=1}^{|A|} p(d_{s_i}|w_s) \log \frac{p(d_{s_i}|w_s)}{p(d_{t_i}|w_t)} \ , \tag{7}$$

$$\mathbf{sim}(w_s, w_t) = \exp\big(-\beta D(\theta_{w_s} \| \theta_{w_t})\big) \ . \tag{8}$$

In this formulation, A is the alignment vector in Eq. 1, θ_{w_s} and θ_{w_t} are the source and target word models respectively, and β is a free parameter that controls the weight of translations in the resulting translation language model. Extracting related words in the target language in addition to the translation(s) is a benefit of using comparable corpora for the CLIR task. But, according to the noisy structure of comparable corpora, the reliability of extracted target words decrease as their rank increase. An appropriate solution to control the effect of lower rank words is proposed in [17]. This issue is addressed in our approach by considering parameter β in the similarity function by analogy with the *word disambiguation* work [5]. Based on the defined similarity function in Eq. 8, the target words are ranked for each source word. To calculate the above similarity score, we need to estimate source and target word models, i.e. $\{p(d_{s_i}|w_s)\}_{i=1}^{|A|}$ and $\{p(d_{t_i}|w_t)\}_{i=1}^{|A|}$ respectively. Using Bayes' Rule, we have:

$$p(d_i|w) = \frac{p(w|d_i)p(d_i)}{\sum_{j=1}^{|A|} p(w|d_j)p(d_j)} , \tag{9}$$

in which w is w_s or w_t, and similarly d_i is source or target document. So, we need to calculate $\{p(w|d_i)\}_{i=1}^{|A|}$ and prior probabilities of documents. The basic way for estimating document language model is maximum likelihood estimator which results in $p_{ml}(w|d_i) = \frac{\mathbf{tf}(w,d_i)}{|d_i|}$. This estimation is not appropriate for calculating KL-Divergence in Eq. 8 as we might have $p(w_s|d_{s_i}) > 0$ while $p(w_t|d_{t_i}) = 0$ which cause $\log \infty$. To resolve this problem, we adopt smoothing methods in estimating laguage models of the target documents. Two commonly used smoothing methods are *Jelinek-Mercer Method* and *Dirichlet Prior Smoothing* [22].

The next step is to derive the prior probabilities for documents. Intuitively alignments with higher similarity scores are more trustable. To take alignment qualities into account, the prior probability of a document (d_{s_i}/d_{t_i}) is estimated based on the probability of the alignment containing that document (a_i). Alignment probabilities are calculated by normalizing the alignment similarity scores:

$$p(d_{s_i}|w_s) = \frac{p(w_s|d_{s_i})p(a_i)}{\sum_{j=1}^{|A|} p(w_s|d_{s_j})p(a_j)}, \quad p(a_i) = \frac{sim(d_{s_i}, d_{t_i})}{\sum_{j=1}^{|A|} sim(d_{s_j}, d_{t_j})} . \tag{10}$$

$p(d_{t_i}|w_t)$ is calculated in a similar way. With this estimation of source and target word models, similarity scores of Eq. 8 can be calculated for each pair of source-target words.

4 Experiments

In this section, experiments concerning cross-language information retrieval between English and Persian languages are described. The English words are stemmed, but the Persian words are not, due to the lack of a good stemmer for this language. Also, stop words are removed. All experiments are done using the Lemur toolkit [2]. Also, only the title of queries are used in all experiments and for each experiment, Mean Average Precision (MAP) and Precision at 10 documents (Prec@10) are reported.

Table 1. Baseline results

Data Set	Monolingual(KL-divergence)		Cross-lingual(Dictionary)	
	MAP	Prec@10	MAP(% Mono-IR)	Prec@10
Ham'08	0.4231	0.6460	0.1161 (27.44%)	0.2060
Ham'09	0.3710	0.6020	0.1041 (28.05%)	0.2286
INFILE	0.4196	0.5047	0.0961 (22.90 %)	0.1547

Table 2. CLIR Performance using LM-based translation language model

Data Set	k-fold Results		Optimal Results				
	MAP(% Mono-IR)	Prec@10	β	λ	Num	MAP(% Mono-IR)	Prec@10
Ham'08	0.1743 (41.19%)	0.304	16	0.7	8	0.1833 (43.32%)	0.3160
Ham'09	0.1097 (29.56%)	0.206	8	0.5	2	0.1301 (35.06%)	0.2360
INFILE	0.2193 (52.26 %)	0.3309	10	0.8	4	0.2222 (52.95 %)	0.3310

4.1 Data Sets

The comparable corpus which is used as the translation resource for the following experiments is UTPECC (University of Tehran Persian-English Comparable Corpora) version 2.0 [14]. It has been constructed from 5-year BBC news in English and 5-year Hamshahri news in Persian. UTPECC includes 14979 alignments which aligns 10724 BBC news with 5544 Hamshahri news.

For Cross-language evaluation purpose, two document collections are used: (1) Hamshahri collection consisting of 166,774 documents in Persian with two sets of CLEF topics, 551-600 and 601-650 in Persian and English (2) INFILE collection (CLEF 2009 INFILE track) consisting of 100,000 documents from Agence France Press (AFP) newswire stories in English evaluated with topics $\{101, ..., 150\} - \{104, 108, 110, 112, 119, 124, 134, 147\}$ in English with their translations in Persian [8].

4.2 Baseline Results

For evaluating cross-language results, we first provide monolingual retrieval results for each test collection. KL-divergence retrieval model is used for monolingual runs with Dirichlet prior smoothing, in which μ is set to 1000. Table 1 shows the results. The CLIR performance using *FarsiDic* machine-readable dictionary [1] is also reported in Table 1. These results are obtained using retrieval model in Eq. 5 assuming uniform probabilities for all translations of each word in the dictionary. The CLIR performance using dictionary is lower in English-Persian, compared to many reported results in other languages (which is above 50% in most cases). The reasons should be investigated in the future.

Table 3. Statistics on coverage of translation resources

Query Set	# of Queries	# of Words	Translation Resource			
			Dictionary		Comparable Corpus	
			# of Translated Queries	# of Translated Words	# of Translated Queries	# of Translated Words
Ham'08	50	149	30	136	42	140
Ham'09	50	148	25	124	41	138
INFILE	42	115	21	94	34	107

4.3 Evaluating the Proposed Approach

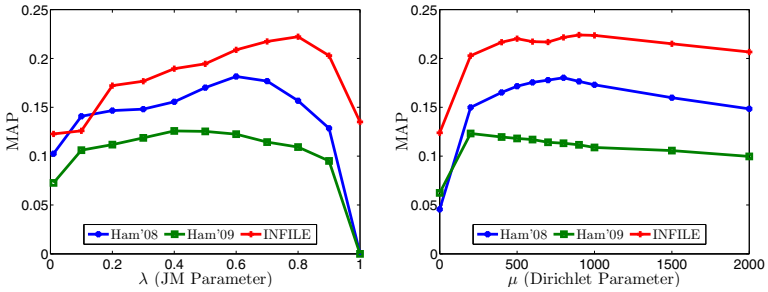
In this section, we investigate extracting translation model using our proposed *language modeling* method and the CLIR performance using the extracted translation model. In these experiments, we use *Maximum Likelihood* estimator for estimating source document language models. However, language models of the target documents should be smoothed which is done by *JM* smoothing method. Thus similarity function in Eq. 8 has two parameters that need tuning: β and λ (JM smoothing parameter). To investigate the effectiveness of the proposed approach, we play down the impact of tuning by employing k-fold cross validation method. The reported MAP is the average of MAP values of k test folds. In this experiment, 4 top translations for each source word are used for building the translation language model. Table 2 shows the k-fold results. We use 5-fold cross-validation for evaluation using *Ham'08* and *Ham'09* query sets, and 3-fold for *INFILE* query set as it has 42 queries. Using comparable corpus, we achieved an effectiveness of 41.19% of monolingual performance for *Ham'08*, which is 50.12% improvement over dictionary-based CLIR. Tuning the parameters leads to further improvements in the performance. Table 2 also shows the optimal results which are obtained by exhaustive parameter search. We also tune the number of selected translations for each source word which is reported in the "Num" column of Table 2. In this case, the CLIR performance is 43.32% of monolingual performance for the *Ham'08* data set.

To illustrate how comparable corpus can help improve performance, we explore two criteria: coverage and quality of the translation resource. To compare the coverage of the translation resources, we count the query words that are translated using each resource. These statistics are reported in the Table 3. The table also includes the number of queries that are completely translated using each translation resource (i.e. all terms of the query are translated). As shown in the table, using comparable corpus, we can translate more number of words compared to using dictionary. Some of the OOV words such as 'wimbledon' can be translated appropriately by the comparable corpus, but not the dictionary.

For translation quality comparison, we select the queries that are completely translated by both resources, and compare the two resources according to the MAP values of this derived query set. Among 50 queries of *Ham'08* query set, 29 queries are selected according to the mentioned criteria. Comparing the CLIR performance for these queries shows that using one translation resource is not

Table 4. CLIR performance using merged resources

Data Set	α	MAP(% Mono-IR)	Prec@10
Ham'08	0.5	0.2137 (50.50 %)	0.3380
Ham'09	0.6	0.1620 (43.66 %)	0.3160
INFILE	0.5	0.2448 (58.34 %)	0.3286



(a) variation of JM parameter (b) variation of Dirichlet parameter

Fig. 1. CLIR performance using different smoothing parameters

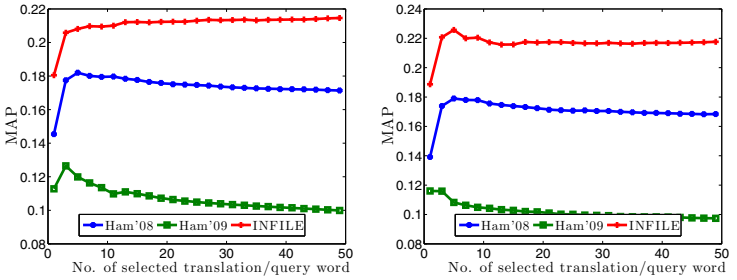
superior to the other in all cases. But, on average comparable corpus outperforms the dictionary (0.1959 versus 0.1754), which shows the advantage of comparable corpus in extracting words that co-occur with the translations.

We also study the effectiveness of combining dictionary and comparable corpus translations. For this purpose, uniform probabilities are assigned to the dictionary translations of each source word. Two translation resources are linearly combined: $p_{comb}(w_t|w_s) = \alpha p_{dic}(w_t|w_s) + (1-\alpha)p_{cc}(w_t|w_s)$. The translations from comparable corpus which yield the optimal results in Table 2 are combined with dictionary translations. We tune the combination parameter α and the best CLIR performance using the combined translation language model is reported in Table 4. Using combined translation language model outperforms using each resource independently.

4.4 Effect of Smoothing on Translation Quality

Language models of both source and target documents in Eq. 9 can be smoothed. When smoothing is not used for the source documents, the summation in Eq. 8 can be calculated only for the alignments that contain the source word, which reduces the calculation time. Therefore, to investigate the influence of smoothing on the quality of extracted translation model, we only smooth the language models of the target documents by considering both *JM* and *Dirichlet prior* smoothing strategies.

We vary the *JM* smoothing parameter (λ) and measure the CLIR performance using the extracted translation model in each case. For cross-language retrieval, 4 top translations of each source query word are selected from the extracted



(a) Smoothing with JM method (b) Smoothing with Dirichlet prior

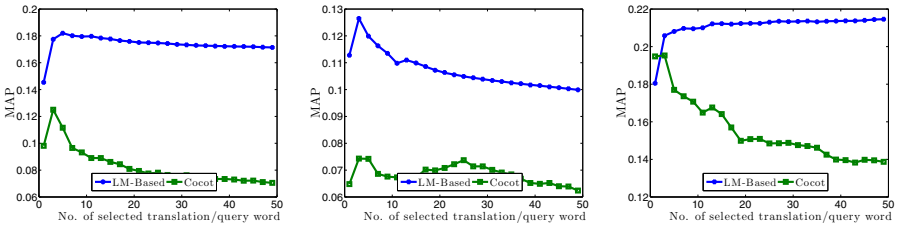
Fig. 2. Sensitivity of MAP to the number of words used for translation

translations and β is set to 10. Fig. 1(a) reports the effect of *JM* smoothing parameter on the CLIR performance. Results from Fig. 1(a) clearly demonstrate that λ is better to be higher than 0.3 as we need to make the probabilities of high entropy words less different for all documents. We get an acceptable value for MAP for a wide range of λ. Moreover, for $\lambda \geq 0.8$, the MAP drops sharply because in these cases differences of documents will be ignored. In a similar way, we study the effect of *Dirichlet prior* smoothing parameter. Fig. 1(b) shows the CLIR performance according to the variation of *Dirichlet prior* smoothing parameter using the previous configuration (number of selected translations = 4 and $\beta = 10$). The results from Fig. 1(b) confirms that language models of the target documents should be smoothed but not substantially. Our experiments also show that the optimal value for μ is about the average document length for each data set.

In addition, the CLIR performance is sensitive to the number of translations selected for each source query word. So, we investigate the sensitivity of our approach to this parameter. The results are shown in Fig.2. In these experiments, we set $\lambda = 0.6$, $\mu = 800$ and $\beta = 10$. With increasing the number of selected translations, at first the MAP curve rises to a level and then stays there. The curve demonstrates that our weighting approach is appropriate and does not allow noise words to pull down the MAP. In addition, if we select few numbers of extracted translations, MAP decreases as we lose some good translations.

4.5 Comparison with Other Approaches

In this section, we compare our proposed method for estimating translation model with *Cocot* [19], *Frequency Correlation-Based* [20] and *Spider* [18] approaches. In Fig. 3, we compare the CLIR performance using *Cocot* approach and our proposed approach. Higher MAP values demonstrate that our approach extracts better translations with more appropriate weights. In a similar way, Fig. 4 depicts the CLIR performance using *FC-Based* approach for translation extraction compared to our proposed method. The *FC-Based* method does not consider existence of alignments that share a same document, while it is addressed in *Cocot* by creating hyper documents and in our approach by involving alignment

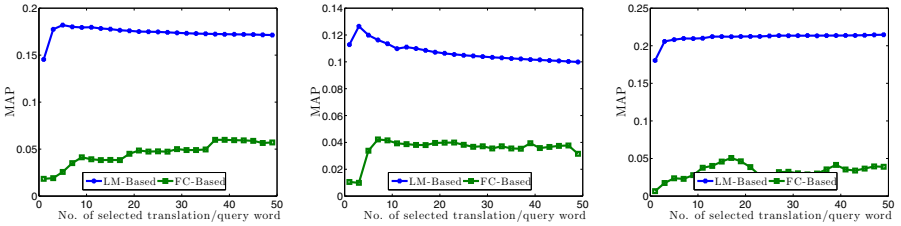


(a) Hamshahri'08 collection (b) Hamshahri'09 collection (c) Infile collection

Fig. 3. CLIR performance using Cocot & LM-Based approach

Table 5. Performance of translation extraction approaches

Data Set	MAP				Prec@10			
	LM-Based	Cocot	FC-Based	Spider	LM-Based	Cocot	FC-Based	Spider
Ham'08	0.1743	0.1250	0.0599	0.0148	0.304	0.2260	0.1156	0.0596
Ham'09	0.1097	0.0743	0.0422	0.0115	0.206	0.1280	0.0723	0.0449
INFILE	0.2193	0.1953	0.0507	0.0307	0.3309	0.2762	0.0839	0.0457



(a) Hamshahri'08 collection (b) Hamshahri'09 collection (c) Infile collection

Fig. 4. CLIR perf. using FC-Based & LM-Based approach

Table 6. Translation quality

Translation Direction	MAP			
	LM-Based	Cocot	FC-Based	Spider
English-Persian	0.0759	0.0576	0.0467	0.0098
Persian-English	0.1141	0.0996	0.0803	0.0245

probabilities. This might cause low performance of *FC-Based* approach. In addition, the diagrams in Fig. 3 and Fig. 4 reflect the sensitivity of MAP measure to the number of selected translations for each source word. From Fig. 3 and Fig. 4, it is clear that the proposed method shows more robust behavior in terms of the selected number of translations.

Table 5 summarizes the results of *Cocot*, *FC-Based*, *Spider* and *LM-Based* approaches. The reported results for *Cocot*, *FC-Based* and *Spider* approaches are

the best that can be achieved through different numbers of selected translations for each source word. For *LM-Based* approach, the k-fold results are mentioned. The results indicate that our approach is better in finding translation or related words. Improvements over other approaches are statistically significant with a 95% confidence according to the Wilcoxon signed-rank test for MAP measure.

To compare the quality of extracted translations from the comparable corpus using different approaches , we use dictionary translations as reference. We measure the MAP of the top 5 extracted translations using each approach based on dictionary translations. The results are shown in Table 6. As the table shows, our approach improves the translation quality in both directions over the previous approaches.

5 Conclusions and Future Work

In this paper, we proposed and evaluated a language modeling approach for extracting translation language models. The focus of our paper is to provide a more practical, effective way for estimating the translation language models. By several experiments, we demonstrate that the proposed method can improve the translation quality as well as the CLIR performance with easier parameter tuning in comparison with similar approaches.

There are many possible directions to extend this work. In this work, we study a simple way of estimating language models, proposing other ways for generating word models will be helpful. Investigating how translation knowledge from other resources can be integrated in the process of extracting translations from comparable corpora is another future research direction. This could be an alternative for the current solution which is combining translations extracted from each resource separately.

References

1. Farsi dictionary, `http://www.farsidic.com/`
2. Lemur toolkit, `http://www.lemurproject.org/`
3. AbduI-Rauf, S., Schwenk, H.: On the use of comparable corpora to improve SMT performance. In: Proceedings of EACL 2009, pp. 16–23. Association for Computational Linguistics, Stroudsburg (2009)
4. Chiao, Y.C., Zweigenbaum, P.: Looking for candidate translational equivalents in specialized, comparable corpora. In: Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002, vol. 2, pp. 1–5. Association for Computational Linguistics, Stroudsburg (2002)
5. Dagan, I., Lee, L., Pereira, F.: Similarity-based methods for word sense disambiguation. In: Proceedings of ACL 1998, pp. 56–63. Association for Computational Linguistics, Stroudsburg (1997)
6. Garera, N., Callison-Burch, C., Yarowsky, D.: Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, pp. 129–137. Association for Computational Linguistics, Stroudsburg (2009)

7. Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL 2004. Association for Computational Linguistics, Stroudsburg (2004)
8. Hashemi, H.B.: Using Comparable Corpora for Persian-English Cross Language Information Retrieval. Master's thesis, University of Tehran (2011)
9. Hazem, A., Morin, E.: Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul (2012)
10. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, pp. 644–652. Association for Computational Linguistics, Stroudsburg (2010)
11. Li, B., Gaussier, E., Aizawa, A.: Clustering comparable corpora for bilingual lexicon extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, HLT 2011, vol. 2, pp. 473–478. Association for Computational Linguistics, Stroudsburg (2011)
12. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. 31(4), 477–504 (2005)
13. Nie, J.Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
14. Rahimi, Z., Shakery, A.: Topic based creation of a persian-english comparable corpus. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 458–469. Springer, Heidelberg (2011)
15. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of ACL 1995, pp. 320–322. Association for Computational Linguistics, Stroudsburg (1995)
16. Sadat, F., Yoshikawa, M., Uemura, S.: Enhancing cross-language information retrieval by an automatic acquisition of bilingual terminology from comparable corpora. In: Proceedings of ACM SIGIR 2003, pp. 397–398. ACM, New York (2003)
17. Shakery, A., Zhai, C.: Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. Information Retrieval, 1–29 (2012)
18. Sheridan, P., Ballerini, J.P.: Experiments in multilingual information retrieval using the spider system. In: Proceedings of ACM SIGIR 1996, pp. 58–65. ACM, New York (1996)
19. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., Keskustalo, H.: Creating and exploiting a comparable corpus in cross-language information retrieval. ACM Trans. Inf. Syst. 25(1) (February 2007)
20. Tao, T., Zhai, C.: Mining comparable bilingual text corpora for cross-language information integration. In: Proceedings of the ACM SIGKDD, KDD 2005, pp. 691–696. ACM, New York (2005)
21. Vulić, I., Moens, M.F.: Detecting highly confident word translations from comparable corpora without any prior knowledge. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012, pp. 449–459. Association for Computational Linguistics, Stroudsburg (2012)
22. Zhai, C.: Statistical language models for information retrieval: A critical review. Foundations and Trends in Information Retrieval 2(3), 137–213 (2008)