Creating a Persian-English Comparable Corpus

Homa Baradaran Hashemi, Azadeh Shakery, and Heshaam Faili

School of Electrical and Computer Engineering,
College of Engineering,
University of Tehran
H.B.Hashemi@ece.ut.ac.ir,{Shakery,HFaili}@ut.ac.ir

Abstract. Multilingual corpora are valuable resources for cross-language information retrieval and are available in many language pairs. However the Persian language does not have rich multilingual resources due to some of its special features and difficulties in constructing the corpora. In this study, we build a Persian-English comparable corpus from two independent news collections: BBC News in English and Hamshahri news in Persian. We use the similarity of the document topics and their publication dates to align the documents in these sets. We tried several alternatives for constructing the comparable corpora and assessed the quality of the corpora using different criteria. Evaluation results show the high quality of the aligned documents and using the Persian-English comparable corpus for extracting translation knowledge seems promising.

1 Introduction and Related Work

The fast growth of the World Wide Web and the availability of information in different languages have attracted much attention in research on cross-language information retrieval (CLIR). One of the main issues in CLIR is where to obtain the translation knowledge [15]. Multilingual corpora, either parallel or comparable, are widely used for this purpose and are available in many language pairs. Comparable corpora are generally obtained from news articles [14, 5, 22, 2], novels [7], available research corpora such as CLEF or TREC collections [23, 4, 21] or by crawling the web [24, 20, 26]. On the other hand, parallel corpora are usually obtained from official multilingual documents such as United Nations articles [6] and EU documents [12], multilingual websites [18] or news services [27]. However, Persian, as a widely spoken language in the Middle East, does not have rich resources due to some of its special features and difficulties in constructing corpora [8, 1].

The current available Persian corpora are either monolingual and built for special purposes or not big enough for translation purposes. For example, Hamshahri corpus [1] is a monolingual corpus for evaluating Persian information retrieval systems and Bijankhan corpus [3] is a Persian tagged corpus for natural language processing. Current available Persian-English corpora are the Miangah's English-Persian parallel corpus [13] consisting of 4,860,000 words, Tehran

English-Persian parallel corpus composed of 612,086 bilingual sentences extracted from movie subtitles in conversational text domain [16], and Karimi's comparable Persian-English corpus [9] consisting of 1100 loosely translated BBC News documents. Since the translation knowledge is usually extracted statistically from the multilingual corpora, the available corpora are not adequate, being either too small or in a special domain.

In this work, we build a Persian-English comparable corpus using Persian articles of Hamshahri newspaper[1] and English articles of BBC News[2]. The total of around 53,000 English documents are aligned with 190,000 Persian documents resulting a comparable corpus of more than 7,500 document pairs. Many studies on exploiting comparable corpora for CLIR assume that comparable corpora are easily obtainable from news articles in different language aligned by date [25, 26, 2]. Although this assumption may be true in some languages, this is not the case in Persian. News from the news agencies that produce daily news articles in both English and Persian are not appropriate, since usually English articles are very short and in most cases are translated summaries of Persian articles and besides their archives are not available online. On the other hand, articles from different news agencies are not easily aligned by date, since in many cases the same event is published in different dates. Thus we were made to choose two distinct collections in different origins and align the articles to obtain the comparable corpus using their publication dates and content similarity scores.

We follow the general procedure proposed in [4], [23] to construct our comparable corpus. Talvensaari et. el, in [23] present a method to build comparable corpora from two collections in different languages and different origins. In their work, they extract the best query keys from documents of one collection using the Relative Average Term Frequency (RATF) formula [17]. The keys are translated into the language of the other collection using a dictionary-based query translation program. The translated queries are run against the target collection and documents are aligned based on the similarity scores and their publication dates. Their method is tested on a Swedish newswire collection and an American newspaper collection. Their approach is most closely related to Braschler et. el, [4] which introduced a method to align documents in different languages by using dates, subject codes, common proper nouns, numbers and a small dictionary.

However, our method for constructing the corpus is different from [4], [23] in many details, somewhat because of the language differences and available linguistic resources. Talvensaari et. al [23], use TWOL lemmatizer [10] to lemmatize inflected Swedish document words and to decompose compound words, while we do not do any preprocessing on our Persian collection. Another difference is in translating the keywords. Talvensaari et. al used UTACLIR, a dictionary-based query translation program, which uses query structuring to disambiguate translation alternatives and a fuzzy string matching technique to transform words not found in the dictionary [23] to translate document keys, while we use a simple

[1] www.hamshahrionline.ir
[2] www.bbc.co.uk

dictionary and Google translator[3] to translate missing words from dictionary. The existence of many non-translated English words in our work suggests that we should use transliteration. Using Google transliteration[4] is shown to be beneficial in our experiments. Moreover, Talvensaari et. al, use a lot of heuristics for their alignments and for setting the parameters and thresholds, while we try to be more general and do not include these heuristics in our work.

In our experiments, we evaluated several methods of creating comparable corpora and the experiment results show that (1) Using top-k translation alternatives of a word from dictionary can improve the quality of comparable corpus over using all translations of a word. (2) Using transliteration besides using dictionary and machine translation improves accuracy. (3) Using feedback retrieval model helps.

The rest of the paper is organized as follows. We explain the details of constructing our comparable corpus in section 2. We present the experiment results in section 3 and finally bring the conclusions and future work of our study in section 4.

2 Constructing the Comparable Corpus

To construct the comparable corpus, we start with two independent news collections, one in English and another in Persian, and try to align the documents in these collections. In our comparable corpus, we want the aligned documents to be as similar as possible. Intuitively, two documents are similar if their corresponding keywords - words that best describe the topic of the document - are close to each other. The publication date is another factor for finding good alignments. Documents with similar content which are published on the same date are most probably talking about the same event. Based on these intuitions, we follow these steps to align the documents: Extract the keywords of the documents in the source language, translate the keywords to the target language and run the translated queries against the target collection. We then align the documents based on their similarity scores and dates. In the rest of this section, we present the details of the method for constructing and evaluating the comparable corpus.

2.1 Query Construction and Translation

In order to construct query words for each source document, we applied the RATF formula [17].

$$RATF(k) = (cf_k/df_k) \times 10^3/ln(df_k + SP)^p, \quad (1)$$

In this formula, df_k and cf_k are document frequency and collection frequency of word k respectively. SP is a collection dependent scaling parameter to penalize rare words and p is a power parameter. We set these parameters to their best values reported in Talvensaari et. el, [23] which were SP = 1800 and p = 3. In order to construct queries which represent the source documents, we first sort

[3] http://translate.google.com/
[4] www.google.com/transliterate/persian

the terms of each document in decreasing order of their frequencies within the document. Equal frequency keys are then sorted according to their RATF values. Finally, 30 top ranked keys are selected as the query which represents the document. Since not all the source language keys can be translated to the target language, we chose to select a slightly large number of keywords to represent each document. To translate selected keywords, we use an English-Persian Dictionary with more than 50,000 entries. Since there are lots of Out Of Vocabulary (OOV) words such as proper nouns, we use Google's machine translation system to translate the words not found in our dictionary. We then use Google transliteration system to translate the words which are still not translated, including some proper nouns and stemmed words.

After creating the query in the target language, we use a retrieval model to rank the target documents based on their similarities to the query. We use these similarity scores along with publication dates to align the documents.

2.2 Document Alignment

In order to create the alignment pairs, we use two basic criteria: the similarity scores of the documents and their publication dates. Intuitively, if an English document and a Persian document have a high similarity score and are published at the same date, they are likely talking about a related event. On the other hand, if the similarity score is low or the publication date is distant, the document pair doesn't seem to be a good match. We apply a combination of three different score thresholds ($\theta_1 < \theta_2 < \theta_3$) to search for suitable document alignments. The values of thresholds are percentiles, e.g. $\theta = 60$ means that the score is greater than 60% of all the similarity scores in the runs. If there are n source documents and for each source document, r target documents are tested, then $n \times r$ scores should be considered to calculate the thresholds [23]. In our experiments, we set r to 50. The steps of document alignment are as follows. Considering an English document, we first search in its top r most similar target documents to find Persian documents which are published in the same day and also their similarity scores are greater that θ_1. If we couldn't find any alignment pair, the threshold increases to θ_2 and we search for target documents in a period of four days. Finally, if we still have not found any matching document, the score threshold is increased to θ_3 and a period of fourteen days is searched.

2.3 Comparable Corpora Evaluation

The main criterion for evaluating a comparable corpus is the quality of the alignments. In our experiments, we assessed the quality of alignments using a five-level relevance scale. The relevance scale is gained from [4]. The five levels of relevance are:

(1) Same story. Two documents are exactly about the same event.
(2) Related story. Two documents deal with same events but in somewhat different viewpoints. (e.g. one document may be part of the other document)
(3) Shared aspect. The documents cover two related events. (e.g. events in the same location or about same people)

(4) Common terminology. The similarity between the events is slight, but a considerable amount of terminology is shared.
(5) Unrelated. There is no apparent relation between the documents.

Which classes to be considered as good alignments depends on the intended application. For example, Braschler et. el, [4] considered classes (1) through (4) to be helpful for extracting good terms in CLIR systems. In our experiments, we count classes (1) through (3) as good alignments. Thus a high quality corpus is expected to have most of its alignments in levels (1), (2) and (3) and not many alignments in levels (4) and (5).

We also used other criteria for further evaluation of the corpus, for example the ability to extract meaningful word associations from the documents and the size of high quality discovered alignments.

In order to extract word associations from the comparable corpus, we used the method proposed in Talvensaari et. al, [24]. The intuition of this method is to use co-occurrence of words in the alignments to extract word associations. The algorithm first calculates a weight w_{ik} for each word s_i in document d_k as:

$$w_{ik} = \begin{cases} 0 & \text{if } tf_{ik} = 0 \\ (0.5 + 0.5 \times \frac{tf_{ik}}{Maxtf_k}) \times ln(\frac{NT}{dl_k}) & \text{otherwise} \end{cases} \quad (2)$$

where tf_{ik} is the frequency of s_i in document d_k , $Maxtf_k$ the largest term frequency in d_k and dl_k is the number of unique words in the document. NT can be either the number of unique words in the collection or its approximation. This $tf.idf$ modification is adopted from Sheridan and Ballerini [21] who also used it in similarity thesaurus calculation. The weight of a target word t_j in a set of ranked target documents D is calculated as:

$$W_j = \sum_{r=1}^{|D|} \frac{w_{jr}}{ln(r+1)}, \quad (3)$$

where D is the set of target documents aligned with a source document containing t_j. The documents in D are ranked based on their alignment scores. Less similar documents, which appear lower in the list, are trusted less for translation and their weights are penalized. This penalization is achieved by $ln(r+1)$ in the denominator.

Finally, the similarity score between a source word s_i and a target word t_j can be calculated as

$$sim(s_i, t_j) = \frac{\sum_{\langle d_k, D\rangle \in A} w_{ik} \times W_j}{\parallel s_i \parallel \times((1-\alpha) + \alpha \times \frac{\|t_j\|}{\overline{\|T\|}})}, \quad (4)$$

in which w_{ik} is the weight of source word s_i in the source document d_k, W_j is the weight of target word t_j in the set of target documents D which are aligned with the source document d_k, A is the set of all alignments, $\parallel s_i \parallel$ is s_i's norm vector, $\overline{\parallel T \parallel}$ is the mean of the target term vector lengths, and α is a constant between 0 and 1 (we chose $\alpha = 0.2$ same as [24]). In this formula, Pivoted vector normalization scheme is employed to compensate long feature vectors.

Table 1. Statistics on the English and Persian Document Collections

Collection	# of Docs.	Time Span	Avg. Doc. Length	# Unique Terms
BBC	53697	Jan. 2002-Dec. 2006	461	141819
Hamshahri	191440	Jan. 2002-Dec. 2006	527	528864

3 Experiments and Results

In this section, we report our experiments on creating the Persian-English comparable corpora and the analysis of their qualities. In our experiments, we have used the Lemur toolkit[5] as our retrieval system. We used Porter stemmer for stemming the English words and Inquery's stopword list (418 words).

3.1 Document Collections

We have used news articles in Persian and English as our documents in Persian-English comparable corpora. Our English collection is composed of news articles published in BBC News and our Persian collection includes the news articles of Hamshahri newspaper. We have used five years of news articles, dated from Jan. 2002 to Dec. 2006. The BBC articles are crawled from the BBC News website and preprocessed to clean the web pages, and also to omit local news of United Kingdom, which will not be aligned with any Persian news article. The Hamshahri articles are extracted from Hamshahri corpus[6] which consists of all of the Hamshahri news articles published between 1996 and 2007. The details of the collections are given in Table 1.

3.2 Creating and Evaluating the Comparable Corpus

To construct the comparable corpus, we first extract the keywords of each document in the source language and translate the keywords to the target language (see section 2.1). These translations are considered as queries in the target language and are run against the target collection to retrieve a ranked list of related documents. The results are processed using the method explained in section 2.2 and tested with different document relevance score thresholds to create the document alignments and thus the comparable corpus.

We have experimented with different alternatives for (1) source language keyword translation, and (2) retrieval models. In order to compare the quality of different alignments, corresponding to different comparable corpora, we manually assessed the quality of alignments for one month, Jan. 2002, on a five-level relevance scale (see section 2.3). Our evaluation results show that different alternatives for constructing the comparable corpora result in corpora with very different qualities.

In our first set of experiments, we used a simple dictionary to translate the source keywords and used Google translator to translate the query words not found in the dictionary. We used the KL-divergence retrieval model with pseudo relevance feedback as our retrieval model [11]. In our experiments, we set the

[5] http://www.lemurproject.org/
[6] http://ece.ut.ac.ir/dbrg/Hamshahri/

Table 2. Assessed Quality of Alignments in one month, Jan. 2002, of experiments with KL-divergence retrieval model. (a)using all dictionary translations of each word. (b)using top-3 translations of each word

(a)

	# of Alignments	% of Alignments
Class 1	4	11.76
Class 2	4	11.76
Class 3	7	20.58
Class 4	11	32.35
Class 5	8	23.52
Total	34	100

(b)

	# of Alignments	% of Alignments
Class 1	3	6.97
Class 2	17	39.53
Class 3	14	32.55
Class 4	8	18.6
Class 5	1	2.32
Total	43	100

Table 3. Results of top-3 Translations With or Without Transliteration and KL-divergence Retrieval Model

	No Transliteration		Transliteration	
	# of alignments	% of alignments	# of alignments	% of alignments
Class 1	3	6.97	5	9.43
Class 2	17	39.53	24	45.28
Class 3	14	32.55	14	26.41
Class 4	8	18.6	8	15.09
Class 5	1	2.32	2	3.77
Total	43	100	53	100

score thresholds to $\theta_1 = 60, \theta_2 = 80$ and $\theta_3 = 95$. Table 2 (a) shows the assessed quality of alignments in this set of experiments using all available translations of a word in the dictionary. As the table shows, roughly 45% of the assessed alignments are about related events and more than half of the alignments are in classes (4) and (5) having little or no similarity.

Different words in our dictionary have different number of translations and this may bias our translated queries. In our second set of experiments, we just used the top three translations of each word in the dictionary to construct the query in the target language. Table 2 (b) shows the results of alignments in the second set of experiments. We can see that almost 79% of the alignments are about related events. This indicates, a big quality improvement by using top-3 translations of each word.

Using the dictionary accompanied with machine translator, there are still some source keywords which are not translated. These words are either proper nouns (such as Euroland, Brinkema, Moussaoui, Toiba and Belkheir) or stem words (such as Sydnei, Melbourn and Athen) which seem to have a high impact in our alignments. In our next set of experiments, we tried to transliterate those words which still are not translated. Table 3 shows the results of adding Google transliteration for missing words. We repeat the results with no transliteration for easier comparison. As can be seen from the table using transliteration can bring in better alignment pairs.

Table 4. Results of top-3 Translation With or Without Transliteration and Okapi Retrieval Model

	No Transliteration		Transliteration	
	# of alignments	% of alignments	# of alignments	% of alignments
Class 1	11	13.58	13	14.94
Class 2	46	56.79	51	58.62
Class 3	20	24.69	19	21.83
Class 4	4	4.93	4	4.59
Class 5	0	0	0	0
Total	81	100	87	100

Table 5. Statistics on the Constructed Comparable Corpus

# of alignments	7580
# of unique English (BBC) documents	4267
# of unique Persian (Hamshahri) documents	3488
# of alignments in same day	1838
# of alignments in four day period	2786
# of alignments in fourteen day period	2956

In our next set of experiments, we used Okapi with pseudo relevance feedback as our retrieval model [19]. We used the top 3 words of the dictionary for translation. Table 4 shows the results. Since our main goal is to find as many high quality document pairs as possible, we compare the size of the aligned corpora as well as their quality. As can be seen from tables 3 and 4, by using transliteration, the number of discovered alignments increases and interestingly, the newly discovered ones are mostly distributed in classes (1) and (2) which shows that most of the new aligned pairs are highly or fairly related.

This set of experiments has the best results among all our alignment experiments and we use this set of aligned documents as our comparable corpus. This comparable corpus consists of 7,580 document alignments. Using that specified thresholds, 8% of the 53,697 source documents are aligned. Table 5 shows some statistics about our created comparable corpus. Since the source and target documents are very different, the relatively low number of alignments was expected. Moreover, the number of alignments can be increased with lowering the thresholds, but this can also affect the quality of the comparable corpus. Table 6 shows the result with $\theta_1 = 0, \theta_2 = 60$ and $\theta_3 = 95$. As the table shows lowering the thresholds bring the 43.7% more aligned documents but percentage of high quality alignments drops from 95.4% to 86.4%. Our comparable corpus contains 76% of all the high quality alignments with very small number of low quality ones. The quality of alignments is crucial when extracting translation knowledge from the corpus.

3.3 Extracting Word Associations

As another criterion to examine the quality of our comparable collection, we tried to extract word associations from the corpus and assess the quality of obtained associations (see section 2.3). Naturally, the higher the quality of the comparable

Table 6. Assessment of Alignment Quality for Two Different Sets of Score Thresholds

	$\theta_1 = 60, \theta_2 = 80, \theta_3 = 95$		$\theta_1 = 0, \theta_2 = 60, \theta_3 = 95$	
	# of alignments	% of alignments	# of alignments	% of alignments
Class 1	13	14.94	16	12.8
Class 2	51	58.62	62	49.6
Class 3	19	21.83	30	24
Class 4	4	4.59	14	11.2
Class 5	0	0	3	2.4
Total	87	100	125	100

Table 7. Top Term Similarities

English Word	Persian Word	Google Translation of Persian Word	Score	English Word	Persian Word	Google Translation of Persian Word	Score
iraqi	عراق	Iraq	104.91	korea	شمالی	North	78.86
korea	کره	Korea	93.73	market	بازار	Market	78.41
elect	انتخابات	Election	89.73	price	قیمت	Price	78.07
nuclear	هسته	Nucleus	85.02	economi	اقتصاد	Economy	77.77
champion	مدال	Medal	82.63	econom	اقتصادی	Economic	74.80
weapon	عراق	Iraq	82.15	oil	نفت	Oil	74.24
cancer	سرطان	Cancer	80.73	attack	حمله	Attack	73.02

Table 8. Word Associations for Four English Words

English Word	Persian Word	Google Translation of Persian Word	Score	English Word	Persian Word	Google Translation of Persian Word	Score
korea	کره	Korea	93.73	cancer	سرطان	Cancer	80.73
	شمالی	North	78.86		بیماری	Disease	52.01
	پیونگ	Pyvng	71.77		بدن	Body	51.26
	ینگ	Yang	71.40		سلول	Cell	41.67
	جنوبی	South	61.35		مبتلا	Suffering	39.97
iraqi	عراق	Iraq	104.91	champion	مدال	Medal	82.63
	صدام	Saddam	95.05		المپیک	Olympics	82.20
	عراقی	Iraqi	82.97		قهرمان	Champion	73.50
	بغداد	Baghdad	82.78		قهرمانی	Championship	72.26
	حسین	Hussein	75.29		مسابقات	Competitions	72.17

corpora, the more precise the word associations will be. Table 7 shows a sample set of top English-Persian associated word pairs extracted from our comparable corpus. We also include the Persian words' Google translations for the readers not familiar with Persian. As can be seen from the table, most of the matched words have a very high quality. We should note that we are showing the stemmed English words in this table and that's why some of the suffixes are missing.

In Table 8, we show the top Persian words aligned with four of the English words. As the table shows, the confidence of matching decreases as we go down the list but the word pairs are still related. This observation suggests that these results can be used in query expansion.

Table 9. Query Translation using Comparable Corpus versus Dictionary

Method	MAP	% of Mono	Prec@5	% of Mono	Prec@10	% of Mono
Mono Baseline	0.42		0.62		0.596	
CC-Top-1	0.111	26.33	0.208	33.54	0.17	28.52
CC-Top-2	0.14	33.30	0.244	39.35	0.232	38.92
CC-Top-5	0.116	27.51	0.216	34.83	0.19	31.87
Dic-Top-1	0.12	28.46	0.212	34.19	0.18	30.20
Dic-Top-3	0.13	30.84	0.192	30.96	0.202	33.89
Dic-Top-5	0.153	36.29	0.224	36.12	0.206	34.56
Dic-all	0.139	32.97	0.2	32.25	0.184	30.87

3.4 Cross-Language Experiments

In the next step of our research, we intend to do cross-language information retrieval using the obtained cross-lingual word associations from the comparable corpus. As the cross-language information retrieval task we focus on the CLIR task of CLEF-2008[7]: Retrieval of Persian documents from topics in English. The document collection for this task contains 166,774 news stories (578MB) that appeared in the Hamshahri newspaper between 1996 and 2002. The queries are 50 topic descriptions in Persian and the English translations of these topics. The Persian queries are used for monolingual retrieval.

In this study, we use the top k associated words in Persian to translate a query word in English with the intuition that these translations are more reliable. We normalize the raw scores to construct translation probabilities and construct the corresponding Persian query language model for each English query. We then rank the documents based on the KL-divergence between the estimated query language models and the document language models. We use monolingual Persian retrieval as a baseline to which we compare the cross-language results. In our monolingual Persian runs, we only use the title field of each Persian query topic as the query words. Since there is not any reliable stemmer for Persian, we did not stem the Persian words. Table 9 shows the results of CC-Top-1, CC-Top-2 and CC-Top-5 translations, where we use the top 1, 2 and top 5 mined associated words from comparable corpora as the translations of each query word.

We also did another run of experiments for CLIR using a dictionary as our translation knowledge. We used the top 1, 3 and top 5 translations of each of the query words from the dictionary to translate the queries. As can be seen from the table, using only comparable corpora and compared to the monolingual baseline, we can achieve up to 33.3% of mean average precision, 39.35% of precision at 5 documents and 38.92% of precision at 10 documents. Using dictionary, we can achieve about 36.29% of mean average precision, 36.12% of precision at 5 and 34.56% of precision at 10 documents. These results show that using only comparable corpora as a translation resource to perform cross-language information retrieval is comparable to using dictionary naively, i.e., constructing the

[7] www.clef-campaign.org

query in the target language by using all translations of each query word in the dictionary.

4 Conclusions and Future Work

In this work, we created a Persian-English comparable corpus, which is, to the best of our knowledge, the first big comparable corpus for Persian and English. We created the corpus from two independent news collections and aligned the documents based on their topic similarities and publication dates. We experimented with several alternatives for constructing the comparable corpora, such as different translation methods and different retrieval models. We assessed the quality of our corpus using different criteria. As the first and most important criterion, we used a five-level relevance scale to manually evaluate the quality of alignments for one month. The evaluation results show that by properly translating the query words and using Okapi with pseudo relevance feedback as the retrieval model, we can come up with a high quality comparable corpus, for which 95% of the assessed matched articles are highly or fairly about related events.

We also tried to extract word associations from the comparable corpus and evaluate the quality of obtained associations. Furthermore, we did cross-language information retrieval using the cross-lingual word associations extracted from the comparable corpus. Experiment results show promising results for extracting translation knowledge from the corpus, although it needs more exploration.

In our future work we are going to focus on CLIR task by improving the quality of extracted word associations. We will try to use the comparable corpus, along with other linguistic resources such as dictionaries, machine translation systems or parallel corpora to improve the CLIR performance. It will also be interesting to use the extracted translation knowledge to improve the quality of the created corpus, by using the extracted word associations as an additional resource to translate source language keywords.

Acknowledgments. This research is partially supported by Iran Telecommunication Research Center (ITRC).

References

1. AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F.: Hamshahri: A standard Persian text collection. Knowledge-Based Systems 22(5), 382–387 (2009)
2. Bekavac, B., Osenova, P., Simov, K., Tadić, M.: Making monolingual corpora comparable: a case study of Bulgarian and Croatian. In: LREC. pp. 1187–1190 (2004)
3. Bijankhan, M.: Role of language corpora in writing grammar: introducing a computer software. Iranian Journal of Linguistics (38), 38–67 (2004)
4. Braschler, M., Schäuble, P.: Multilingual information retrieval based on document alignment techniques. In: ECDL. pp. 183–197 (1998)
5. Collier, N., Kumano, A., Hirakawa, H.: An application of local relevance feedback for building comparable corpora from news article matching. NII J (Natl Inst Inform) 5, 9–23 (2003)

6. Davis, M.W.: On the effective use of large parallel corpora in cross-language text retrieval. Cross-language information retrieval pp. 11–22 (1998)
7. Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H.J., Tufis, D.: Multext-east: parallel and comparable corpora and lexicons for six central and eastern european languages. In: ACL. pp. 315–319 (1998)
8. Ghayoomi, M., Momtazi, S., Bijankhan, M.: A study of corpus development for Persian. International Journal of Asian Language Processing 20(1), 17–33 (2010)
9. Karimi, S.: Machine Transliteration of Proper Names between English and Persian. Ph.D. thesis, RMIT University, Melbourne, Victoria, Australia (2008)
10. Koskenniemi, K.: Two-level morphology: A general computational model for word-form recognition and production. Publications of the Department of General Linguistics, University of Helsinki 11 (1983)
11. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: SIGIR. pp. 111–119 (2001)
12. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: SIGIR. pp. 159–166 (2002)
13. Miangah, T.M.: Constructing a Large-Scale English-Persian Parallel Corpus. Meta: Translators' Journal 54(1), 181–188 (2009)
14. Munteanu, D., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist 31(4), 477–504 (2005)
15. Oard, D., Diekema, A.: Cross-language information retrieval. Annual Review of Information Science and Technology 33, 223–256 (1998)
16. Pilevar, M.T., Feili, H.: PersianSMT: A first attempt to english-persian statistical machine translation. In: JADT (2010)
17. Pirkola, A., Leppanen, E., Järvelin, K.: The RATF formula (Kwoks formula): exploiting average term frequency in cross-language retrieval. Information Research 7(2) (2002)
18. Resnik, P.: Mining the web for bilingual text. In: ACL. pp. 527–534 (1999)
19. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: SIGIR. pp. 232–241 (1994)
20. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: WaCky! Working papers on the Web as Corpus (2006)
21. Sheridan, P., Ballerini, J.P.: Experiments in multilingual information retrieval using the spider system. In: SIGIR. pp. 58–65 (1996)
22. Steinberger, R., Pouliquen, B., Ignat, C.: Navigating multilingual news collections using automatically extracted information. CIT 13(4), 257–264 (2005)
23. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M.: Creating and exploiting a comparable corpus in cross-language information retrieval. TOIS 25(4) (2007)
24. Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., Laurikkala, J.: Focused web crawling in the acquisition of comparable corpora. Information Retrieval 11, 427–445 (2008)
25. Tao, T., Zhai, C.X.: Mining comparable bilingual text corpora for cross-language information integration. In: SIGKDD. pp. 691–696 (2005)
26. Utsuro, T., Horiuchi, T., Chiba, Y., Hamamoto, T.: Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. In: AMTA. pp. 165–176 (2002)
27. Yang, C.C., Li, W., et al.: Building parallel corpora by automatic title alignment using length-based and text-based approaches. Information Processing & Management 40(6), 939–955 (2004)