

Cross-Lingual Semantic Similarity Measure for Comparable Articles

Motaz Saad, David Langlois, and Kamel Smaïli

SMa^rT Group, LORIA,
INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{motaz.saad,david.langlois,kamel.smaili}@loria.fr

Abstract. A measure of similarity is required to find and compare cross-lingual articles concerning a specific topic. This measure can be based on bilingual dictionaries or based on numerical methods such as Latent Semantic Indexing (LSI). In this paper, we use LSI in two ways to retrieve Arabic-English comparable articles. The first way is monolingual: the English article is translated into Arabic and then mapped into the Arabic LSI space; the second way is cross-lingual: Arabic and English documents are mapped into Arabic-English LSI space. Then we compare LSI approaches to the dictionary-based approach on several English-Arabic parallel and comparable corpora. Results indicate that the performance of our cross-lingual LSI approach is competitive to the monolingual approach and even better for some corpora. Moreover, both LSI approaches outperform the dictionary approach.

Keywords: Cross-lingual latent semantic indexing, corpus comparability, cross-lingual information retrieval.

1 Introduction

Comparing cross-lingual articles is a challenging problem for several tasks in natural language processing and especially in machine translation and cross-lingual information retrieval. The comparison can be done in terms of topics, opinions or emotions. In this paper, we focus on how to retrieve comparable articles. A comparable corpus is a collection of articles in multiple languages which are not necessarily translations of each other, but they are related to the same topic. On the other hand, a parallel corpus can be considered as a comparable corpus in which each sentence in the source corpus is aligned to its translation in the target corpus.

There are many methods proposed in literature to compare as well as to retrieve cross-lingual articles. These methods are based on bilingual dictionaries [10,16,19], or on cross-lingual Information retrieval (CL-IR) [7,1,21] or on cross-lingual Latent Semantic Indexing (CL-LSI) [2,11,6,14].

In dictionary-based methods [10,16,19], two cross-lingual documents d_a and d_e are comparable if a maximum of words in d_a are translations of words in

d_e , so a bilingual dictionary can be used to look-up the translation of words in both documents. The drawbacks of this approach are the dependency on bilingual dictionaries which are not always available and the necessity to use morphological analyzers for languages that can be inflected. Moreover, word-to-word translations based on dictionaries can lead to many errors. [19] proposed binary and cosine measures based on multi-WordNet [3] dictionary to compare Wikipedia and news articles. Both binary and cosine measures proposed by [19] require the source-target texts to be represented as vectors of aligned words. Word weight for the binary measure is either 1 or 0 (presence or absence of the word), while it is term frequency for the cosine measure. The similarity of cross-lingual documents is computed as follows: the binary measure counts the words in d_a which are translation of words in d_e and then normalize it by the vector size, whereas the cosine measure computes the cosine similarity between source and target vectors which represent the frequency of the aligned words of d_a and d_e .

In Cross-Lingual Information Retrieval (CL-IR) methods, one can use Machine Translation (MT) systems in order to achieve source and target documents into the same language. Then classical IR tools can be used to identify comparable articles [7,1,21]. Query documents are usually translated into the language of indexed documents. This is because the computational cost of translating queries is far less than the cost of translating all indexed documents. The drawback of this approach is the dependency on MT systems. The performance of MT affects the performance of the IR system. Moreover, the MT system needs to be developed first if it is not available for the desired language.

In Cross-Lingual Latent Semantic Indexing (CL-LSI) methods, documents are described as numerical vectors that are mapped into a new space. Then one can compute the cosine between vectors to measure the similarity between them. The LSI method has already been used in context of CL-IR in [2,11,14]. In their approach, the source document and its translation (the target) are concatenated into one document and then LSI learns links between source and target words or documents. [2] focused their work on Greek-English document retrieval and [11] focused on French-English documents, while [14] computed the similarity of Wikipedia articles in several European languages.

In this work, we focus on CL-IR for English-Arabic document retrieval. In order to avoid using bilingual dictionaries or morphological analyzers or MT systems, we use CL-LSI to compare and retrieve English-Arabic documents. Another advantage of CL-LSI is that it overcomes the problem of vocabulary mismatch between queries and documents. We therefore use the same approach as [11], however, we apply it on Arabic-English articles and [11] used parallel corpus in their work, but we use both parallel and comparable corpus to train CL-LSI.

In this paper, we use LSI in two ways to retrieve Arabic-English comparable documents. We refer to the first way as monolingual: the English article is translated and then mapped into the LSI Arabic space; the second way as cross-lingual: Arabic and English articles are mapped into Arabic-English CL-LSI

space. We also compare these methods to the dictionary-based method proposed by [19] which is described above.

Besides using CL-LSI to retrieve comparable articles, we also use it to measure the “comparability of a corpus”, i.e. to inspect if a target corpus is a translation of a source one and how much they are different from each other. This enables an understanding of how much the source and target texts, in a comparable corpora, are similar to each other. This can be useful for many applications such as cross-lingual lexicon extraction, information extraction, and sentence alignment.

The rest of the paper is organized as follows: corpora and the method are described in Sect. 2, 3, and 4. Results are presented and in Sect. 5. Finally, the conclusion is stated.

2 Corpora

In this section we describe the corpora we used for our experiments. It consists of documents collected from newspapers, United Nations resolutions, talks, movie subtitles and other domains. These corpora are either parallel or comparable. A detailed description of these corpora is provided in the following subsections.

2.1 Parallel Corpora

Table 1 presents the parallel corpora. $|S|$ is the number of sentences, $|W|$ is the number of words, and $|V|$ is the vocabulary size. The table also shows the domain of each corpus. The parallel corpora that we use are: AFP¹, ANN², ASB³ [12], Medar⁴, NIST [15], UN [17], TED⁵ [4], OST⁶ [20] and Tatoeba⁷ [20].

Note that OST is a collection of movie subtitles translated and uploaded by users. So the quality of the translations may vary from a user to another.

As can be noted from Table 1, in all parallel corpora, English texts have more words than Arabic. In contrast, Arabic texts have vocabulary larger than English. The reason is that certain Arabic terms can be agglutinated [13], while English terms are isolated. For instance, the Arabic term *وَسَنُعْطِيهِمْ* *wasanoṭeyhm* translating to “and we will give them” in English, is an example where one Arabic term corresponds to five English words. On the other hand, Arabic has a larger vocabulary because it is morphologically rich [8,18]. For example, the English word “travellers” may correspond to three forms in Arabic: *مُسَافِرُونَ* *mosāferwn* in masculine nominative form, *مُسَافِرِينَ* *mosāferyn* in masculine accusative/genitive form or *مُسَافِرَات* *mosāferāt* in feminine form.

¹ www.afp.com

² www.annahar.com

³ www.assabah.com.tn

⁴ www.medar.info

⁵ www.ted.com

⁶ www.opensubtitles.org

⁷ www.tatoeba.org

Table 1. Parallel Corpora

Corpus	S	W		V	
		English	Arabic	English	Arabic
Newspapers					
AFP	4K	140K	114K	17K	25K
ANN	10K	387K	288K	39K	63K
ASB	4K	187K	139K	21K	34K
Medar	13K	398K	382K	43K	71K
NIST	2K	85K	64K	15K	22K
United Nations Resolutions					
UN	61K	2.8M	2.4M	42K	77K
Talks					
TED	88K	1.9M	1.6M	88K	182K
Movie Subtitles					
OST	2M	31M	22.4M	504K	1.3M
Other					
Tatoeba	1K	17K	13K	4K	6K
Total	2.3M	37M	27.5M	775K	1.8M

2.2 Comparable Corpora

Table 2 shows WIKI and EuroNews comparable corpora, where $|D|$ is the number of articles, $|W|$ is the number of words and $|V|$ is the vocabulary size. Each pair of comparable articles is related to the same topic. WIKI and EuroNews corpora were collected and aligned at article level in [19]. WIKI is collected from Wikipedia website⁸ and EuroNews is collected from EuroNews website.⁹ WIKI articles are edited online by Wikipedia community. There is a hyperlink between articles that are related to the same topic, but each article may be written independently. Therefore, Wikipedia articles are not necessarily translations of each other.

Table 2. Comparable Corpora

	WIKI		EuroNews	
	English	Arabic	English	Arabic
$ D $	40K	40K	34K	34K
$ W $	91.3M	22M	6.8M	5.5M
$ V $	2.8M	1.5M	232K	373K

3 LSI-Based Methods

The LSI method [5] decomposes a term-document matrix X using the the Singular Value Decomposition (SVD) as $X = USV^T$. The matrices U and V^T are

⁸ www.wikipedia.org

⁹ www.euronews.com

the left and right singular vectors respectively, while S is a diagonal matrix of singular values. Each column vector in matrix U maps terms in the corpus into a single concept of semantically related terms that are grouped with similar values in U . The decomposition USV^T has a rank R , where R is the reduced concept dimensionality in LSI.

For our monolingual LSI approach, X is represented as in (1). It is an $m \times n$ matrix that represents a given monolingual corpus which consists of n documents, and m terms. The entries w_{ij} are the *tfidf* weights.

$$X = \begin{matrix} & d_1 & d_2 & \dots & d_n \\ \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{matrix} & \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{pmatrix} \end{matrix} \quad (1)$$

$$X = \begin{matrix} & d_1^u & d_2^u & \dots & d_n^u \\ \begin{matrix} t_1^a \\ t_2^a \\ \vdots \\ t_l^a \\ t_1^e \\ t_2^e \\ \vdots \\ t_m^e \end{matrix} & \begin{pmatrix} w_{11}^a & w_{12}^a & \dots & w_{1n}^a \\ w_{21}^a & w_{22}^a & \dots & w_{2n}^a \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1}^a & w_{l2}^a & \dots & w_{ln}^a \\ w_{11}^e & w_{12}^e & \dots & w_{1n}^e \\ w_{21}^e & w_{22}^e & \dots & w_{2n}^e \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1}^e & w_{m2}^e & \dots & w_{mn}^e \end{pmatrix} \end{matrix} \quad (2)$$

In our cross-lingual LSI approach, X is represented as in (2). Each d_i^u is the concatenation of the Arabic document d_i^a and its corresponding English document d_i^e . Consequently, X represents a bilingual corpus consisting of n cross-lingual documents, l Arabic terms, and m English terms. So X is an $(l + m) \times n$ matrix. X , as represented in (2), can be used to represent parallel or comparable corpora. For a parallel corpus, each d_i^u represents a pair of parallel sentences, while for a comparable corpus, it represents a pair of comparable documents. Term-document matrix as formulated in (2), enables LSI to learn the relationship between terms which are semantically related in the same language or between two languages.

This method helps us to achieve our objective to retrieve comparable articles. We describe this retrieval process in the next section.

4 Experiment Procedure

As outlined in Sect. 1, for a source document in English, our objective is to retrieve the target comparable documents in Arabic. So the source document is compared with all target documents and then the most similar target documents are retrieved. This is done by describing the source and target documents as

bag-of-words, then mapping them into vectors in LSI space and subsequently by comparing these vectors. If the value of cosine similarity between the two vectors is high, we consider these two documents as comparable. All English and Arabic texts are preprocessed by removing punctuation marks.

In the next sections, we describe how LSI matrices are built and how they are used to retrieve comparable articles. Then we compare the results of these two methods.

4.1 Building LSI Matrices

Steps below describe how LSI matrices are built:

1. Split English and Arabic corpora presented in Sect. 2 into training (90%) and testing (10%) subsets.
2. Use Arabic training corpus to create X as in (1). Then apply LSI to obtain USV^T , the monolingual LSI matrix (LSI-AR) as shown in left of the Fig. 1.
3. Use English-Arabic training corpus to create X as in (2). Then apply LSI to obtain USV^T , the cross-lingual LSI matrix (LSI-U) as shown in right of the Fig. 1.

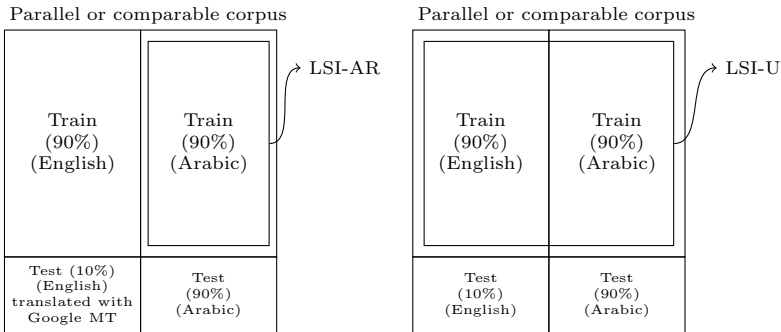


Fig. 1. LSI models

The optimal rank of USV^T in steps 2 and 3 above is chosen experimentally. According to [9], the optimal number of dimensions to perform SVD is in the range [100...500]. We conducted several experiments in order to determine the best rank and we found that the dimension 300 optimizes the similarity for the parallel corpus. So we use the dimension 300 in all our experiments.

4.2 Retrieving Comparable Articles

The test corpus is composed of n pairs of English e_i and Arabic a_j documents (aligned at sentence level in parallel corpus and at the document level in comparable corpus). The goal is then to retrieve the a_i among all the a_j given e_i . The following steps describe the two methods:

LSI-AR:

1. For each a_j , get $a'_j: a'_j = a_j^t U S^{-1}$.
2. Translate each English document e_i into Arabic using Google MT service¹⁰ and get a_{e_i} .
3. For each a_{e_i} , get $a'_{e_i}: a'_{e_i} = a_{e_i}^t U S^{-1}$.
4. For each a'_{e_i} and a'_j , compute $\cos(a'_{e_i}, a'_j)$.

LSI-U:

1. For each a_j , get $a'_j: a'_j = a_j^t U S^{-1}$.
2. For each e_i , get $e'_i: e'_i = e_i^t U S^{-1}$.
3. For each e'_i and a'_j , compute $\cos(e'_i, a'_j)$.

e'_i , a'_{e_i} , and a'_j in the methods above are vectors of the same nature since they have a language independent representation. After these two methods, we can use the cosine values to get the most similar Arabic document to a given English one. For each e_i , we sort a_j in descending order according to the cosine values. e_i and a_j are truly comparable if $i = j$. In other words, for each source document, we have only one relevant document. So in the sorted list of a_j , the condition ($i = j$) is checked in the top-1 (recall at 1 or $R@1$), top-5 (recall at 5 or $R@5$), and top-10 (recall at 10 or $R@10$) lists. The performance measure is defined as the percentage of a_i which are successfully retrieved in $R@1$, $R@5$, $R@10$ lists, among all e_i .

5 Results and Discussion

5.1 Retrieving Parallel Articles

The results of the LSI-AR and LSI-U approaches are presented in Table 3. Results are presented for a random sample of 100 source and target test articles because of the computational cost of doing the experiment on all the test corpus. As shown in Table 3, it is not easy to get a general conclusion about the performance of LSI since it depends on the nature of the corpus and on the desired recall ($R@1$, $R@5$ or $R@10$). For example, for AFP, ASB, TED, UN, and Medar, LSI-U is slightly better than LSI-AR. In contrast, for ANN, NIST, OST and Tatoeba, LSI-AR is better than LSI-U. The performance of LSI-U is equal to, or better than LSI-AR in 6 over 9 of corpora for $R@1$. The average value for ($R@1$) in LSI-AR and LSI-U methods are 0.71 and 0.72 respectively. Moreover, we checked the significance of these differences (McNemar's test), and we found that they are not significantly different. Therefore, both approaches obtain mostly similar performance. In addition, we recall that the LSI-U does not require a MT system. Therefore, we can affirm that the LSI-U is competitive compared to LSI-AR.

¹⁰ translate.google.com

Table 3. LSI results for parallel corpora

Corpus	Method	$R@1$	$R@5$	$R@10$
Newspapers				
AFP	LSI-AR	0.94	0.96	0.99
	LSI-U	0.97	0.99	0.99
ANN	LSI-AR	0.80	0.91	0.94
	LSI-U	0.82	0.92	0.94
ASB	LSI-AR	0.79	0.90	0.92
	LSI-U	0.85	0.92	0.97
Medar	LSI-AR	0.56	0.76	0.81
	LSI-U	0.61	0.78	0.85
NIST	LSI-AR	0.78	0.87	0.92
	LSI-U	0.71	0.82	0.84
United Nations Resolutions				
UN	LSI-AR	0.97	1.00	1.00
	LSI-U	0.98	0.99	1.00
Talks				
TED	LSI-AR	0.52	0.73	0.82
	LSI-U	0.60	0.83	0.92
Movie Subtitles				
OST	LSI-AR	0.39	0.61	0.72
	LSI-U	0.33	0.76	0.85
Other				
Tatoeba	LSI-AR	0.70	0.85	0.94
	LSI-U	0.61	0.79	0.86

The performance of LSI-AR and LSI-U approaches on OST corpus is poor as expected because of the nature of this corpus. OST is composed of subtitles that are translated by many users as mentioned in Sect. 2.

To investigate the effect of the performance of the MT system on the performance of the LSI-AR, we run an experiment to simulate a perfect MT system. This is done by retrieving an Arabic document by providing the same document as a query. This experiment is done on all corpora and the results in terms of $R@1$ are 1.0 for all corpora. These results reveal the lack of robustness of LSI-AR according to the MT system's performance.

We compare our method with the dictionary-based method that was proposed by [19] on the union of AFP and ANN corpora. Results are presented in Table 4 where the dictionary-based method is denoted as DICT.

As shown in the table, both LSI methods achieve better results than DICT, except for $R@10$ which are slightly worse than DICT. It can be concluded that this method is better than DICT since it does not need any dictionary nor morphological analysis and it is language independent.

Table 4. Recall results the union of AFP and ANN corpora

Method	$R@1$	$R@5$	$R@10$
DICT	0.49	0.81	1.0
LSI-AR	0.87	0.95	0.96
LSI-U	0.86	0.96	0.98

5.2 Retrieving Comparable Articles

For comparable corpora, the same experimental protocol is applied. Table 5 shows the performance of recall of the LSI-U method on EuroNews and WIKI comparable corpora. As shown in the table, the performance of the LSI-U on EuroNews corpus is better than WIKI corpus.

Table 5. Testing LSI-U on comparable corpora

Corpus	$R@1$	$R@5$	$R@10$
WIKI	0.42	0.84	0.94
EuroNews	0.84	0.99	1.0

This could be due to the fact that EuroNews articles being mostly translations of each other [19], while Wikipedia articles are not necessarily translations of each other as mentioned in Sect. 2.

From Tables 5 and 3, it can be noted that LSI-U can retrieve the target information at document level and sentence level respectively with almost same performance. The evidence for that is, for parallel corpora, AFP, ANN, and ASB, 0.97, 0.83, and 0.84 $R@1$ was achieved respectively and for EuroNews comparable corpus, 0.84 $R@1$ was achieved.

5.3 Comparing Corpora

We take advantage of the used method in order to study the comparability of some supposed comparable corpora such as WIKI and EuroNews. We do that by computing the average cosine, $\text{avg}(\cos)$, for all pair articles of the test parts of these corpora. So for each corpus, the LSI-U matrix is built from the training part and used to compute the $\text{avg}(\cos)$ for the test part. This experiment is done on BEST, EuroNews and WIKI corpora. BEST is the union of AFP, ASB and UN parallel corpora. These corpora are chosen because they have the best recall performance as shown in Table 3. Statistics on comparability are presented in Table 6.

The average similarity proposes to corroborate the fact that for parallel corpus, we get better recall results than by using the other corpora. In other words, the score for BEST which is a parallel corpus aligned at sentence level is better than the one for WIKI which is considered as a real comparable corpus. For EuroNews (near parallel), which is composed of translated articles, the results are better than for WIKI, but lower than for BEST.

Table 6. Statistics on comparability

Corpus	BEST	EuroNews	WIKI
avg(cos)	0.53	0.46	0.23

6 Conclusion

In this paper we described a method which permits to measure comparability between corpora. This method is based on LSI, which we used in two ways: monolingual (LSI-AR) and cross-lingual (LSI-U). The first method needs to use a machine translation system in order to compare two vectors of the same type of data, whereas the second method merges the training data of both languages and in the test step the comparison is then done on two vectors of the same type since they contain the representation of two cross-lingual documents.

We applied this method on English-Arabic documents. The method allows us to identify comparable articles extracted from a variety of corpora. The measure we proposed has shown its feasibility since it enables distinguishing of parallel corpora from strongly comparable corpora such as Euronews and also from the weakly comparable corpora such as WIKI. The feasibility of the method has been illustrated in this paper since it has been tested on 9 different corpora. Some of them are largely used by the community and others are less popular but more difficult such as OST. The best results have been achieved for AFP corpus and the worst for OST.

In future work we will use this method in order to retrieve comparable articles from the social media to collect and build parallel corpora for languages which are under-resourced. The method developed in this paper will be expanded and adapted in order to compare the cross-lingual corpora in terms of opinions and emotions.

References

1. Aljlal, M., Frieder, O., Grossman, D.: On Arabic-English Cross-Language Information Retrieval: Machine Translation Approach. In: Machine Readable Dictionaries and Machine Translation, ACM Tenth Conference on Information and Knowledge Management (CIKM), pp. 295–302. ACM Press (2002)
2. Berry, M.W., Young, P.G.: Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities* 29(6), 413–429 (1995)
3. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: 6th Global WordNet Conference (GWC 2012), pp. 64–71 (2012)
4. Cettolo, M., Girardi, C., Federico, M.: Wit³: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), Trento, Italy, pp. 261–268 (May 2012)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407 (1990)
6. Dumais, S.: Lsa and information retrieval: Getting back to basics. In: Handbook of Latent Semantic Analysis, pp. 293–321 (2007)

7. Fujii, A., Ishikawa, T.: Applying machine translation to two-stage cross-language information retrieval. In: White, J.S. (ed.) AMTA 2000. LNCS (LNAI), vol. 1934, pp. 13–24. Springer, Heidelberg (2000), http://dx.doi.org/10.1007/3-540-39965-8_2
8. Habash, N.: Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3(1), 1–187 (2010)
9. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. *Discourse Processes* 25(2-3), 259–284 (1998)
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*, pp. 644–652. Association for Computational Linguistics (2010)
11. Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Grefenstette, G. (ed.) *Cross-Language Information Retrieval*. The Springer International Series on Information Retrieval, pp. 51–62. Springer, US (1998)
12. Ma, X., Zakhary, D.: Arabic newswire english translation collection. Linguistic Data Consortium, Philadelphia (2009)
13. Meftouh, K., Laskri, M.T., Smaïli, K.: Modeling Arabic Language using statistical methods. *Arabian Journal for Science and Engineering* 35(2C), 69–82 (2010)
14. Muhic, A., Rupnik, J., Skraba, P.: Cross-lingual document similarity. In: *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI)*, pp. 387–392 (June 2012)
15. NIST, M.I.G.: NIST 2008/2009 open machine translation (OpenMT) evaluation. Linguistic Data Consortium, Philadelphia (2010)
16. Otero, P., López, I., Cilenis, S., de Compostela, S.: Measuring comparability of multilingual corpora extracted from wikipedia. In: *Iberian Cross-Language Natural Language Processings Tasks (ICL)*, p. 8 (2011)
17. Rafalovitch, A., Dale, R.: United nations general assembly resolutions: A six-language parallel corpus. In: *Proceedings of the MT Summit XII*, vol. 13, pp. 292–299 (2009)
18. Saad, M.: The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification. Master's thesis, Computer Engineering Dept., Islamic University of Gaza, Palestine (2010)
19. Saad, M., Langlois, D., Smaïli, K.: Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia - Social and Behavioral Sciences* 95, 40–47 (2013), <http://www.sciencedirect.com/science/article/pii/S1877042813041402>, corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC 2013)
20. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair, N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul (2012)
21. Ture, F.: Searching to Translate and Translating to Search: When Information Retrieval Meets Machine Translation. Ph.D. thesis, Graduate School of the University of Maryland, College Park (2013), <http://hdl.handle.net/1903/14502>