Alignment of comparable documents: comparison of similarity measures on French-English-Arabic data

David Langlois, Motaz Saad, Kamel Smaïli

Natural Language Engineering **1** (1): 000–000. Printed in the United Kingdom

Alignment of comparable documents: comparison of similarity measures on French-English-Arabic data

D. LANGLOIS[a] , M. SAAD[b], K. SMAILI[a]
[a] *SMa*^{r}*T* 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
[b] Islamic University of Gaza, Department of Computer Sciences

(*Received 27 April 2018*)

Abstract

The objective, in this article, is to address the issue of the comparability of documents, which are extracted from different sources and written in different languages. These documents are not necessarily translations of each other. This material is referred as multilingual comparable corpora, These language resources are useful for multilingual natural language processing applications, especially for low-resourced language pairs. In this paper, we collect different data in Arabic, English and French. Two corpora are built by using available hyper links for Wikipedia and Euronews. Euronews is an aligned multilingual (Arabic, English and French) corpus of 34k documents collected from Euronews website. A more challenging issue is to build comparable corpus from two different and independent medias having two distinct editorial lines such as British Broadcasting Corporation (BBC) and Al Jazeera (JSC). To build such corpus, we propose to use the Cross-Lingual Latent Semantic approach. For this purpose, documents have been harvested from BBC and JSC web sites for each month of the years 2012 and 2013. The comparability is calculated for each Arabic-English couple of documents of each month. This automatic task is then validated by hand. This led to a multilingual (Arabic-English) aligned corpus of 305 pairs of documents (233k English words and 137k Arabic words). In addition A study is presented in this paper to analyze the performance of three methods of the literature allowing to measure the comparability of documents on the multilingual reference corpora. A recall at rank 1 of 50.16 per cent is achieved with the Cross-lingual LSI approach for BBC-JSC test corpus, while the dictionary-based method reaches a recall of only 35.41 per cent.

1 Introduction

Multilingual texts (parallel or comparable) are useful in several Natural Language Processing applications such as bilingual lexicon extraction (Li and Gaussier, 2010), cross-lingual information retrieval (Knoth, Zilka, and Zdrahal, 2011) and machine translation (Delpech, 2011). A parallel corpus is a collection of aligned sentences, which are translations of each other. Parallel corpora are acquired using human translators, but this is time-consuming and requires a lot of human efforts.

In addition, these resources are not available for all the language pairs. The huge quantity of data on Internet has increased the interest for collecting multilingual corpora related to the same subject. These text resources are known as comparable corpora. Below, we give an example of two short documents, extracted from Wikipedia in English and French, which are an illustration of what is called comparable texts.

Barack Hussein Obama II born August 4, 1961 is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

Barack Hussein Obama II, né le 4 août 1961 à Honolulu (Hawaï), est un homme d'État américain. Il est le 44e et actuel président des États-Unis, élu pour un premier mandat le 4 novembre 2008 et réélu pour un second le 6 novembre 2012.

These two comparable documents are not translation of each other, but they are related to the same topic. In addition, they embed few fragments of sentences which are translation of each other. The idea consisting in using the web to get comparable corpora is very attractive, but the issue is how to align multilingual documents sharing the same topic, or the same news? In other words, from a list of source documents, how to associate each document to its corresponding one from a list of target documents.

Producing comparable corpora has several utilities in Natural Language Processing. Among them, we can mention the improvement of machine translation systems (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009). This material could be used also to create parallel corpora for under-resourced languages (Wołk and Marasek, 2014). Comparable corpora can be used to extract classical bilingual lexicon (Morin and Prochasson, 2011) or transliteration lexicon for under-resourced languages (Abidi and Smaïli, 2011). Our purpose in this article is to lay the groundwork for an automatic press review. In other words, we would like to compare the content of two documents produced by two different news agencies. The documents may be different at the language level or at the level of the editorial lines of the news agencies. For instance, in AMIS (Access Multilingual Information opinionS)[1], a Chist-Era project in which we are involved in, and aims to compare the content of two videos in terms of opinions. The supported languages are Arabic, French and English.

To achieve the previous objectives, one should make the corpora comparable. Therefore, in this paper, we investigate the task of aligning comparable documents. Given a multilingual corpus which is not aligned, the objective is to automatically align source and target documents which are related to the same context. In this article, we tackle the comparability of documents in three languages: Arabic, English

[1] `http://deustotechlife.deusto.es/amis`

and French. There are two kinds of methods are tested in this paper to reach the objective of aligning documents; the first kind is based on lexical information and the second relies on latent information. We compare these methods on three different multilingual corpora; one is provided by the evaluation campaign: BUCC 2015 (Sharoff, Zweigenbaum, and Rapp, 2015). This corpus is extracted from Wikipedia, the experiments consist in aligning French and English documents. The second one; is more suitable for our objective, since it concerns the press. This corpus has been built in a previous work and concerns documents of Euronews (Saad, Langlois, and Smaïli, 2013). For this second corpus, experiments concern the alignments of the three languages Arabic, French and English. The third corpus is more challenging since, the objective is to align two different medias with two different editorial lines in different languages. The concerned medias are BBC and Al Jazeera. It is remarkable that building a comparable corpus from BBC and Al Jazeera is more difficult, since inter-lingual hyper links are missing, contrary to Wikipedia or Euronews.

The rest of this paper is structured as follows. In Section 2, we give a description of some related works about alignment of comparable articles. The three similarity measures we experiment are described in Section 3. In Section 4, we describe the data sets extracted from Wikipedia and Euronews, and in Section 5 we present the method we used to build the BBC-Al Jazeera corpus. Then, we give the results of the different approaches evaluated on the three data sets (Section 6.1). Finally, we propose a discussion and a conclusion.

2 Related works

In (Fung and Cheung, 2004), the authors proposed three levels for non-parallel corpora. These levels are *noisy-parallel*, *comparable* and *quasi-comparable* corpora. Texts in *noisy-parallel* corpora have many parallel sentences roughly in the same order. Texts in *comparable* corpora have topic aligned documents, which are not necessarily translations of each other. *Quasi-comparable* corpora contain bilingual documents that are not necessarily related to the same topic.

Most of the researchers are interested in comparable corpora because they can be used to extract parallel texts for different purposes. This interest continues to receive increasing attention from the research community. For example, ACCURAT[2] (Pinnis et al., 2012; Skadina et al., 2012) is a research project dedicated to finding methods and techniques to overcome the problem of lacking linguistic data for under-resourced languages and narrow domains. For that, the project proposes approaches to align and extract lexical resources. The ultimate objective is to exploit comparable data to improve the quality of machine translation for under-resourced languages. Other researchers and projects (for example, the "TTC: Terminology Extraction, Translation Tools and Comparable Corpora" project[3]) also considered comparable corpora to improve the quality of machine translation (Smith, Quirk,

[2] `www.accurat-project.eu`
[3] `http://www.ttc-project.eu/`

and Toutanova, 2010; Abdul-Rauf and Schwenk, 2011) and to extract bilingual lexicons (Li and Gaussier, 2010).

The objective of some other researchers is to improve the overall comparability of the corpus (Li, 2012) without retrieving specific alignments between documents, but our objective is to build inter-lingual aligned document pairs; in other words, (Li, 2012) focused on corpus level comparability measure while we focus in this work on document level comparability measure. The methods to align comparable documents are numerous. The issue of aligning comparable multilingual documents could be considered such as an information retrieval (IR) task. Cross-Lingual Information Retrieval (CL-IR) is a special case of IR, where the language of the query is different from the language of the documents (Ballesteros and Croft, 1996). In this case, the CL-IR system should unify the language of queries and documents by using for instance a Machine Translation (MT) system (Aljlayl, Frieder, and Grossman, 2002; Ture, 2013; Hieber and Riezler, 2015). The drawback of this approach is the dependence on the MT system, which affects the performance of the IR system. Moreover, the MT system needs to be developed first if it is not available.

Other different approaches are proposed in this topic. For instance in (Ion, Ceauşu, and Irimia, 2011), the authors proposed an Expectation-Maximization algorithm to find the aligned documents. This is achieved by an iterative process which stops when the words in the aligned documents match at best the content of a translation table. This method requires a parallel corpus for training the translation table. In (Vulić and Moens, 2015), the authors presented a method which tries to show the relationship between word embeddings text representation and information retrieval methods. The work showed, among other points, how to model the cross-lingual semantic embedding space without the use of parallel data or bilingual dictionaries.

In (Oshikiri, Fukui, and Shimodaira, 2016), the authors proposed a method called Cross-Lingual Eigenword which differs from the current trend: word embedding based on neural networks such as in (Vulić and Moens, 2015). The authors, instead of using the skip-gram model, extended the spectral Eigenwords (Dhillon, Foster, and Ungar, 2015) to cross-lingual settings with sentence-alignment. Even if Cross-Lingual Latent Semantic Indexing has not been dedicated to word embedding, the authors have used it as a baseline method. This method achieves competitive results on a English-Spanish word translation task. Unfortunately it has not been tested on the alignment of comparable corpus, which is our objective in this article.

3 Experimented alignment methods

In this section, we present some alignment methods of the state-of-the-art, and we experiment them in Section 6. We present the hapax-based method and a dictionary-based method which have been used in BUCC 2015 shared task campaign (Sharoff et al., 2015). This campaign asked the competitors to retrieve comparable documents from a Wikipedia corpus. We also experimented Cross-Lingual Latent Semantic Indexing (CL-LSI) method on the corpora yet used in (Saad, Langlois,

and Smaïli, 2014; Saad, 2015) and on the Wikipedia corpus from the BUCC2015 shared task.

3.1 Hapax

Morin et al. in (Morin, Hazem, Boudin, and Clouet, 2015) proposed a method based only on the lexical content of documents. The method is based on the matching of the hapax words (which occur once) in both source and target documents. For example, this can be the case for *url*, entity names, Hashtags, etc. Then, the more the documents have common hapax words, the more they are close to each other. In the context of this paper, this method is called HAPAX.

3.2 Dictionary-based methods

In dictionary-based methods, two cross-lingual documents d_s and d_t are comparable if most of words in d_s are translations of words in d_t. A bilingual dictionary has to be used to look-up the translations of words in both documents. The similarity can be measured as follows (Li and Gaussier, 2010; Otero and López, 2011):

$$sim(d_s, d_t) = \frac{|w_s \rightarrow w_t| + |w_t \rightarrow w_s|}{|d_s| + |d_t|} \quad (1)$$

where $|w_s \rightarrow w_t|$ is the number of source words (w_s) that are in the bilingual dictionary and have translations in the target document (d_t), and $|w_t \rightarrow w_s|$ is the number of target words (w_t) that are in the bilingual dictionary and have translations in the source document (d_s). $|d_s| + |d_t|$ is the number of words of the source and the target documents that are also in the bilingual dictionary.

Several similarity measures based on a bilingual dictionary have been proposed. Some of them consider the similarity at the corpus level (Li and Gaussier, 2010), while others consider it at the document level, then aggregate similarities of documents in the corpus (Otero and López, 2011).

In our experiments, we used Li & Gaussier measure (Li and Gaussier, 2010); in addition, we take into account common words between the source and target documents even if they do not occur in the bilingual dictionary. For example, if the source document contains *Obama*, and the target document as well, this would increase the similarity. In the context of this article, we call this method *DB* for Dictionary-Based. By this way, the actual used formula is:

$$sim(d_s, d_t) = \frac{|w_s \rightarrow w_t| + |w_t \rightarrow w_s| + 2c_{st}}{|d_s| + |d_t|} \quad (2)$$

where c_{st} is the number of identical (common) words that appear in d_s and in d_t. c_{st} excludes words that are yet taken into account by Equation 1.

The drawbacks of the dictionary based approach are the dependency on a bilingual dictionary, which is not always available, and the necessity to use morphological analyzers for inflected languages. Moreover, word-to-word dictionary translations

without considering the context can lead to many errors because of the polysemy, and because the text is considered only as a bag of independent words.

To avoid to use a bilingual dictionary built by hand, it is possible to automatically build it. This approach was used in (Etchegoyhen and Azpeitia, 2016), which defines a measure based on Jaccard distance between the source and the target documents. As the measure requires that both the documents are in the same language, the target document is translated into the source language by using a translation table obtained by IBM models (Brown, Pietra, Pietra, and Mercer, 1993).

3.3 Cross-Lingual Latent Semantic Indexing (CL-LSI) methods

Document similarity can be estimated at term or semantic level (Harispe, Ranwez, Janaqi, and Montmain, 2015). Generally, semantic similarity is a measure which quantifies the likeness of documents based on the latent semantic of the contents. Semantically related terms are usually referred as *concepts*. Semantic similarity can be measured based on a predefined ontology, which specifies the distance between concepts, or can be measured using statistical methods which find correlations between terms and contexts in a text corpus. Numerous methods in this latter category exist in literature; in (Vulić and Moens, 2014), a list of cross-lingual word similarity is given. One of these methods, which we use in this article, is Latent Semantic Indexing (LSI).

Documents and words in LSI are projected into a reduced space using Singular Value Decomposition. Similar words and documents are mapped closer to each other in the LSI space. More precisely, the documents are represented by a term-document matrix in which each cell contains the frequency of a word into a document. This sparse matrix is rewritten into a term-to-document and a document-to-term matrices, and these matrices are reduced into a more little space by Singular Value Decomposition. Then, it is possible to project every unseen document into the LSI space and to represent the document by a numerical vector. The similarity of two documents is defined by the cosine-based distance between their LSI vectors. The method is parameterized by the size of the LSI space, which is called *number of Topics*. The LSI approach is extended to the Cross-Lingual domain (CL-LSI) by considering a document as the concatenation between source and target documents (Littman, Dumais, and Landauer, 1998). For more details see (Saad, 2015).

The authors in (Blei, Ng, and Jordan, 2003) extended LSI to a probabilistic version called Latent Dirichlet Allocation (LDA). In LDA, documents are represented as random mixtures of latent topics, these topics are probabilistic distributions over words. Each document can be perceived as a mixture of different topics, and every topic is characterized by a distribution over words. LDA is useful for modeling topics that are mentioned in a corpus while LSI is useful to map similar documents and words in a corpus into a reduced feature space (model concepts) (Cui, Liu, Tan, Shi, Song, Gao, Qu, and Tong, 2011). Among all latent semantic works, we use LSI for cross-lingual similarity measure because the aim is to map similar documents and words across languages closer to each other into a reduced feature space. In other words, we aim to model *concepts* rather than *topics*.

4 Data sets

To evaluate the different approaches, we decided to test on two kinds of corpora. The first one is made up of Wikipedia data and is provided by the BUCC 2015 evaluation campaign. We used it because this corpus is used by the research community. But, our focus is news data. Therefore we used also two other corpora coming from news agencies. The first one is made up of Euronews website content and has been collected in a previous work (Saad et al., 2013). The multilingual document pairs from Euronews were easy to collect because there are hyper links between documents about the same news. The second corpus uses two different news agencies: BBC for English, and Al Jazeera for Arabic. Given that BBC news articles and Al Jazeera ones are not connected by hyper links, the task of collecting them was not so simple. We describe in Section 5 the method we use to align English BBC news and Arabic Al Jazeera articles.

4.1 Wikipedia corpus

The Wikipedia data used in our experiments, consists of pairs of articles in French and English provided by BUCC 2015. Table 1 gives statistics about this corpus. In this table $|D|$ is the number of documents in the corpus, $|W|$ is the number of words in the corpus, and $\bar{W}$ is the average number of words per document.

Wikipedia corpus (228k documents pairs) is divided into two parts. We use more than 113k French-English couples for training the method based on CL-LSI, and 114k for test. We did not use a development corpus for Wikipedia because we did not tune parameters on this data. Each French document is compared to each English document of the test corpus. To overcome the issue of the complexity of comparison, we decided to sample the source (French) corpus by selecting 1,003 representative documents. Then, each document of the source sample has been compared to every article of the target corpus (114,802 articles).

Table 1. Wikipedia comparable corpus characteristics

	English	French
$\|D\|$	228k	228k
$\|W\|$	185M	113M
$\bar{W}$	810	495

4.2 Euronews comparable corpus

Euronews data was collected in a previous work (Saad et al., 2013). Euronews is a multilingual news TV channel available in many European languages as well as in Arabic. For the three languages and for the development and the test corpus, 3,345 documents have

been used, while 27,752 documents served for the training. Details are given in Table 2. For an explanation of the notations used in this table, see Section 4.1.

Table 2. Euronews comparable corpus characteristics

	English	French	Arabic
$\|D\|$	34k	34k	34k
$\|W\|$	6.8M	6.9M	5.5M
$\bar{W}$	198	200	161

5 BBC News - Al Jazeera corpus: aligning comparable documents from different sources

In a previous work, we collected comparable corpus by aligning two different news sources: BBC and Al Jazeera news (Saad, 2015). The challenge is less straightforward than for Wikipedia or Euronews because hyper links between comparable articles do not obviously exist for BBC and Al Jazeera. In the next sections we present the alignment method and statistics on the obtained corpus.

5.1 Alignment method

First, we crawl BBC and Al Jazeera websites to collect news articles published in 2012 and 2013 using HTTRACK tool[4]. The data of BBC-Al Jazeera corpus are then split into several sub-corpora. Each sub-corpus is composed of news articles that are published in a given month. Consequently, we obtain 24 sub-corpora for each language as shown in Figure 1. The number of articles in each month is between 70 and 300.

Then, we use the Cross-Lingual Latent Semantic Approach (Littman et al., 1998; Saad, 2015) to measure the similarity between Arabic and English documents. The LSI model is trained on Euronews corpus (Arabic-English part). For each month, we measure the similarity between each Arabic document and the English documents of the same month. This leads to numerous Arabic-English pairs with a similarity score for each one.

We have now to select the correctly aligned documents from these pairs. We decided that two articles are correctly aligned if and only if they share the same exact news. More precisely, two articles focusing, for example, on earthquakes are correctly aligned if this is the same earthquake in the same country, the same day. It was not possible to check manually each pair of documents. Therefore, this handwork is done on the top-15 article pairs retrieved from each of the 24 months (e.g. 360 pairs). Only 305 out of 360 were aligned correctly and have been kept for the further experiments. In the next section, we present statistics about the obtained comparable corpus.

As for Wikipedia and Euronews corpus, we give statistics on this corpus in Table 3.

[4] www.httrack.com

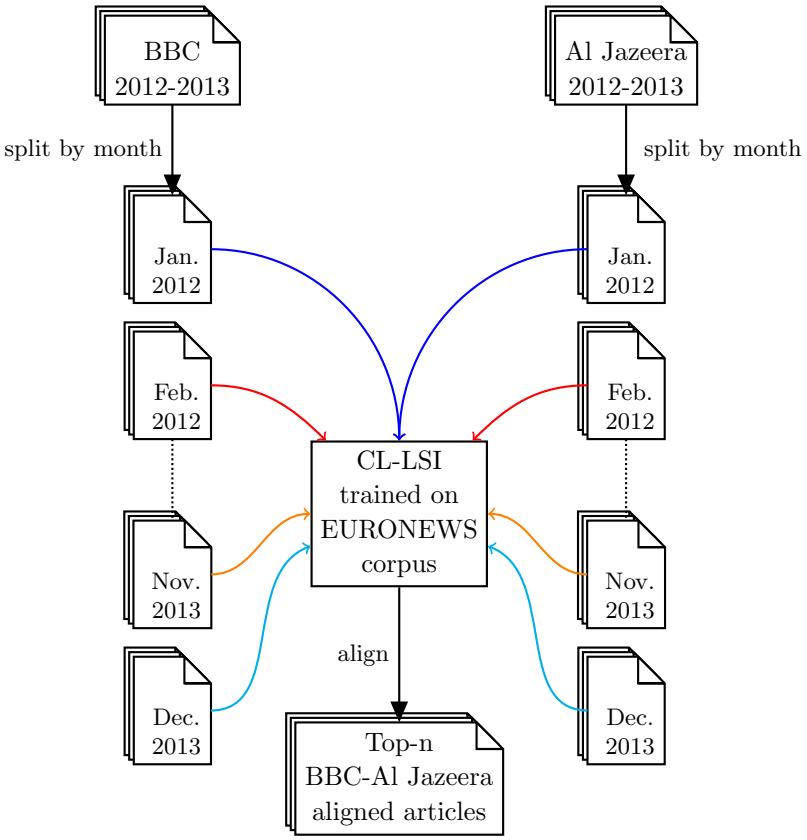

Fig. 1. Automatic alignment of BBC and Al Jazeera news stories

Table 3. BBC-Al Jazeera comparable corpus characteristics

	English	Arabic
$\|D\|$	305	305
$\|W\|$	246k	110k
$\bar{W}$	805	361

5.2 Statistic analysis of the corpus

Figure 2 shows the accuracy of alignment of the top-15 most similar documents of each month. The accuracy of the alignment is defined as the number of cross-lingual articles that are correctly aligned, divided by the total number of articles in the month.

The ranges of similarity values of the top-15 aligned articles for the years 2012 and 2013 are shown in Figure 3. The figure illustrates the minimum and the maximum of similarity values for each month. For 2012, the maximum value is 0.86 and the minimum is 0.45.

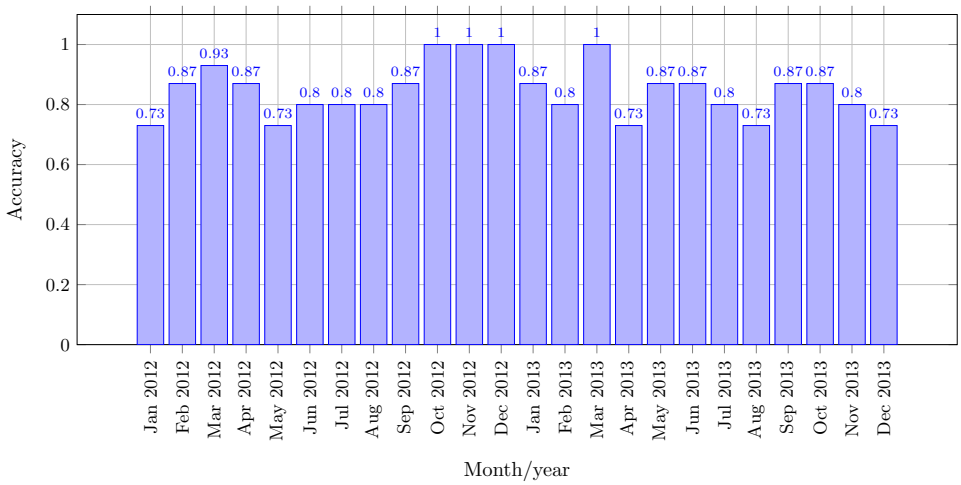

Fig. 2. Accuracy of articles alignment for years 2012 and 2013

For 2013, the maximum value is 0.89 and the minimum is 0.26. It should be noted that the similarity intervals are close to each other for all months in 2012 and 2013, except for January, February, April and May 2013. This is probably due to the nature of the items harvested for each month, where the crawler may miss some items in the collection process. Furthermore, Figure 3 shows that min-max values vary from a month to another. This is why we decided to choose the *top-n* similar articles rather than setting a threshold for the similarity value.

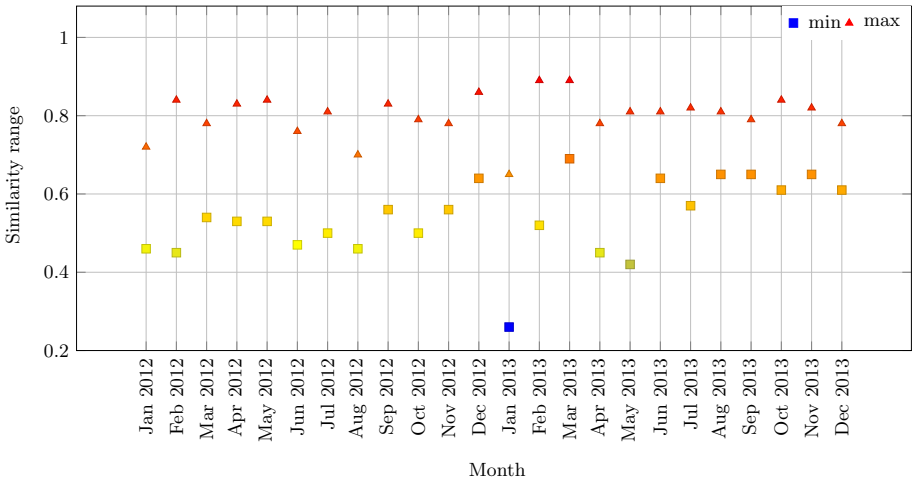

Fig. 3. Similarity ranges of the top-15 similar documents of BBC-Al Jazeera of the years 2012 and 2013

The accuracy of correctly aligned documents is 85 per cent. We carried out more investigations about misaligned articles during the validation process. We found that they are all related to the same topic, but they are not related to the same news story or event. The investigation reveals that some of these articles are misaligned despite their similarity

is high. The reason is that they are related to the same kind of events, but these events happened in different countries. For instance, one of misaligned news articles was related to elections, but the English article was related to elections in Bulgaria while the Arabic article was related to elections in Pakistan. We conducted a search for *elections in Bulgaria* in Al Jazeera collection, but we could not find any news article that is related to this event. We also found that some of these stories are local news that are covered only by one of the two broadcast news: Al Jazeera or BBC. In addition to this, it should also be noted that the crawler sometimes can not harvest all Web pages. That is why, for few months, some news can be found neither in the BBC nor in the collections of Al Jazeera.

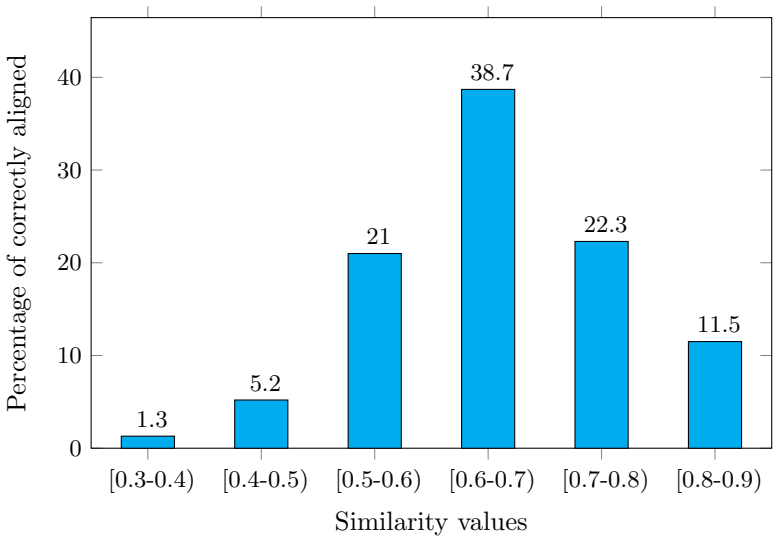

Fig. 4. Distribution of the 305 documents according to the CL-LSI similarity

Figure 4 shows the percentage of correctly aligned articles versus their similarity values. The similarity values in this figure are divided into intervals. The number of correctly aligned articles increases as the similarity value increases, up to the interval $[0.6 - 0.7)$, then it decreases for higher similarity values. The interpretation might be as follows: when the similarity is low, the articles are mostly related to the same topic but not to the same news story. As the similarity increases, the likelihood for the aligned articles to be related to the same news story increases up to a certain value, then it normally decreases again. This is because it is unlikely to find related articles written at the same time by BBC and Al Jazeera, and that have a high similarity value.

6 Experiments on cross-language information retrieval

In this section, we present the results for the two categories of corpus (described in Section 4) as presented in Table 4.

In the experiments, we present the results in terms of recall at rank 1 (@1 per cent) and 5 (@5 per cent) and in terms of Mean Reciprocal Rank (MRR) which is defined below:

$$MRR = \frac{\sum_{d_s \in C} \frac{1}{r(d_t)}}{|C|} \tag{3}$$

Table 4. Language pairs for each corpus

Category	Corpus	Arabic English	French English
non news	**Wikipedia**		X
news	**Euronews**	X	X
news	**English-BBC Arabic-Al Jazeera**	X	

Where d_s is a source document, C is the source corpus, d_t is the target document corresponding to d_s and $r(x)$ is the rank of document x according to the similarity measure.

DB is applied on lemmatised documents because the bilingual dictionary is also lemmatised. For French and English, we used the DELA lemma lists[5]. For Arabic, we used the approach we proposed in a previous work (Saad, 2015).

For the bilingual dictionaries, we used Euradic Sci-Fran[6] for French-English and a dictionary extracted from Multilingual WordNet for Arabic-English. The Euradic Sci-Fran dictionary contains 71k French entries (with 2.1 English translations per entry in average) while the Arabic-English one contains 12k Arabic entries (with 6.5 English translations per entry in average).

The CL-LSI method does not require lemmatisation. But, for verification purposes, we tested CL-LSI on lemmatised and non-lemmatised data. Performance was better with a non-lemmatised corpus. Therefore, in the following experiments, we present only results with such data. Moreover, for this method, we only kept the words occurring more than 3 times.

6.1 Experiments on multi-lingual documents extracted from single source

This section presents the performance of HAPAX, DB and CL-LSI achieved on the three corpora presented in the previous sections.

Table 5 shows the performance of the methods on Wikipedia data.

Since there is no development corpus and because we decided to be in the same test conditions such as in BUCC2015, no tuning has been done. For the LSI method, 300 topics have been empirically chosen (the same value as in (Saad, 2015)).

Table 5 shows that the DB method outperforms the two others whatever the measure. CL-LSI method performance is similar to HAPAX one, except for @5. Performance for the three methods is quite low. These results are explained by the fact that each source document is compared to 114k target documents. the diversity of target documents could mislead while retrieving the relevant document.

Results on Euronews data are given in Tables 6 and 7 for respectively French-English and Arabic-English.

For HAPAX method, we present the results for the French-English pair because this

[5] `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html`

[6] `http://catalog.elra.info/product_info.php?products_id=666`

Table 5. Results on Wikipedia BUCC2015 data

Method	@1 per cent	@5 per cent	MRR
HAPAX	26.22	32.70	0.30
DB	59.12	70.99	0.65
CL-LSI 300 topics	24.00	38.60	0.31

Table 6. Results on Euronews data (French-English)

Method	@1 per cent	@5 per cent	MRR
HAPAX	30.85	53.15	0.41
DB	74.14	87.62	0.80
DB 47 per cent	82.66	94.50	0.88
DB 47 per cent + 0.4 HAPAX	84.33	95.04	0.89
CL-LSI 1000 topics	77.46	95.58	0.85

method is based on the count of the same words that do exist in documents of both languages. For Arabic and English, this method is not applicable because they use different scripts.

The results, in comparison to those given for Wikipedia, are higher. Indeed, this was predictable due to the fact that, for this data set, a source document is compared only to 3,345 articles, which in fact reduces the ambiguity.

The number of topics for CL-LSI has been tuned on a development corpus; the best performance is achieved with 1,000 topics. The results obtained by CL-LSI are better than DB and HAPAX for all the metrics.

For DB method we propose an improvement to boost its results. Upgrading the method consists in selecting the best words of a document in accordance to their tfidf value. In fact, instead of comparing every word of the source document to the words of the target documents, only these representative words are taken into account. In our experiments, these best words are determined by selecting a percentage of the document's vocabulary. This percentage is optimized on a development corpus. The use of a sub-part of document's vocabulary (47 per cent) allows DB to be improved by 11.5 per cent, which outperforms CL-LSI at @1.

We propose also to linearly combine DB with HAPAX in order to improve the results. The weights of this linear combination are also tuned on the development data. Linearly combining DB with HAPAX (with a weight equal to 0.4 for HAPAX) led to an increasing

of 13.7 per cent. This result was unexpected because HAPAX uses limited information in documents.

Table 7. Results on Euronews data (Arabic-English)

Method	@1 per cent	@5 per cent	MRR
DB	37.53	56.06	0.46
DB 71 per cent	46.49	68.07	0.57
CL-LSI 1250 topics	70.00	93.93	0.80

For the data set Arabic-English, CL-LSI obtained the best performance with 1,250 topics. For Arabic, the use of CL-LSI shows a real jump in the performance in comparison to DB method. In fact, CL-LSI exceeds DB by 86.5 per cent. The use of a sub-part of document's vocabulary (71 per cent) allows DB to be improved by 23.9 per cent, but the performance remains lower than CL-LSI. This is probably due to the fact that the Arabic-English dictionary does not cover all words necessary to DB method. When the bilingual dictionary has a larger coverage as the one used for French-English data set, CL-LSI achieves better results than DB but the improvement is not so spectacular. We will investigate this hypothesis in concluding remarks just after experiment sections.

6.2 Experiments on multi-lingual documents extracted from multiple sources

In the previous experiments, tests have been achieved on documents extracted from the the same source. In the following experiment, we would like to test the methods on multi-lingual data extracted from different sources: BBC for English and Al Jazeera for Arabic.

As the BBC-Al Jazeera corpus is very small, 305 documents retrieved automatically and checked by hand (see Section 5), we used this whole corpus as test corpus. The CL-LSI models used here are trained on Euronews (such as in Section 6.1). We tested on the test corpus several values for the number of topics in CL-LSI. The results for each number of topics are given in Table 8. On the 305 documents, 750 topics led to the best performance.

Table 9 shows the performance of the methods on BBC-Al Jazeera data. We can remark that the results of DB method have collapsed. The use of sub-parts of document's vocabulary allows to partly counteract this performance, but CL-LSI remains higher than DB. Even if the performance of CL-LSI decreased also, it could be considered as more robust than DB since the decreasing is not as dramatic.

6.3 Concluding remarks about results

The previous results show that the performance depends strongly on the number of target documents a source article is compared to. For example, for DB (French-English), the performance in terms of @1 is equal to 59.12 for Wikipedia (with more than 100k target articles), and is equal to 75.19 for Euronews (with more than 3k documents). This result seems obvious because the more documents to be compared, the more risk of ambiguity there is between articles.

Table 8. CL-LSI results on BBC-JSC data (test corpus)

# topics	@1 per cent	@5 per cent	MRR	# topics	@1 per cent	@5 per cent	MRR
10	27.87	64.92	0.44	400	47.21	88.20	0.65
20	34.10	74.75	0.52	500	48.52	88.20	0.65
30	38.03	78.36	0.56	**750**	**50.16**	**88.20**	**0.66**
40	39.67	81.31	0.58	1,000	47.87	86.89	0.65
50	42.30	83.28	0.60	1,250	46.56	86.56	0.64
100	46.89	86.56	0.64	1,500	46.23	87.21	0.64
200	47.87	87.54	0.64	1,750	44.92	87.87	0.63
300	49.18	87.21	0.66	2,000	46.23	88.52	0.64

Table 9. Results on BBC-Al Jazeera data (Arabic-English)

Method	@1 per cent	@5 per cent	MRR
DB	16.07	32.46	0.25
DB 40 per cent	35.41	60.98	0.48
CL-LSI 750 topics	50.16	88.20	0.66

The language, also, has a strong impact on performance. DB performance in terms of @1 is 74.14 for Euronews French-English, while it is 37.53 on Euronews Arabic-English. This is certainly due to the weak coverage of the bilingual dictionary since Arabic is strongly agglutinative. Indeed, Table 10 shows that the relative performance for DB is correlated with the dictionary coverage: for the French-English pair, the dictionary has a better coverage than for the Arabic-English pair. A method independent from a bilingual dictionary such as CL-LSI is more robust, despite the fact that its results for Arabic decrease slightly in comparison to French.

Several analyzes may be done on the results of Tables 6 and 7. Concerning the pair of languages French-English of Euronews, where the comparability is measured on a corpus from the same source, we can mention that the best results are provided by DB 47 per

Table 10. Coverage of bilingual dictionaries. DB performance (@1) is given between parentheses.

	Arabic-English	**French-English**
Wikipedia		66.6 per cent (59.12)
Euronews	49.0 per cent (37.53)	74.8 per cent (74.14)
English-BBC Arabic-Al Jazeera	47.5 per cent (16.07)	

cent at @1. This percentage means that only 47 per cent of the most representative words of the document have been kept for the comparison (as explained in Section 3.2). Several rates have been tested, but only the total number and the best one have been reported in Table 6 . The results of LSI method used with 1000 topics are low in comparison to DB, this is probably due to the fact that, texts coming from the same source, it is easier to identify the comparable multilingual documents only by using dictionaries.

HAPAX leads to bad results, since this method is based only on common words between two languages from close languages.

7 Conclusion

In this article, we proposed a method for collecting, retrieving and aligning comparable documents. We used the CL-LSI approach in order to compare documents from two different sources: BBC and Al Jazeera. This allowed to extract Arabic-English couples of documents which have been checked by hand and this led to 305 comparable documents.

Moreover, we studied three similarity measures on a comparable documents retrieval task. We experimented these measures on three corpora strongly different in terms of source (encyclopedia, news) and in terms of size. For the dictionary-based similarity measure, we proposed to use only a selection of the document's vocabulary, stemming on the tfidf value of words. The selection of the best words showed a real improvement for DB. Actually, DB performance is improved by 11.5 per cent on Euronews French-English, and by 23.9 per cent on Euronews Arabic-English. Moreover, DB has very low performance (16.07) in terms of @1 on BBC-Al Jazeera data set, but when the selection of the most representative words is performed, the method obtains better results (35.41). HAPAX achieves, in the majority of experiments, the worst performance in terms of @1, @5 and MRR. Nevertheless, it can improve slightly the performance when it is interpolated with DB.

The results obtained by CL-LSI are better than DB and HAPAX for all the metrics and all the corpora except for Wikipedia. CL-LSI showed its strength when the linguistic resources are missing or has a weak coverage. For the data set BBC Al Jazeera, the DB method, even by selecting the vocabulary, is lower than CL-LSI in terms of performance by 29.4 per cent. Consequently, CL-LSI remains robust even on difficult task as the one concerning the identification of comparable documents from two contrasting information sources such as BBC news and Al Jazeera news. In a future work, we will use the methods described in this paper to collect comparable corpora from social networks because this

media is now a source of information for journalists. Obviously, other issues will rise due to the particularity of this kind of data.

References

Abidi, K. and Smaïli K. 2018. An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings. *In 11th edition of the Language Resources and Evaluation Conference, LREC 2018* Miyazaki, Japan. European Language Resources Association.

Abdul-Rauf, S. and H. Schwenk. 2009. On the use of comparable corpora to improve SMT performance. *In the proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* Athens, Greece, pp. 16–23.

Abdul-Rauf, S. and H. Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4):341–375.

Aljlayl, M., O. Frieder, and D. Grossman. 2002. On Arabic-English Cross-Language Information Retrieval: Machine Translation Approach. *In Proceedings of the International Conference on Information Technology: Coding and Computing*, ITCC '02, pp. 2–, Washington, DC, USA, 2002. IEEE Computer Society.

Ballesteros, L. and B. Croft. 1996. Dictionary methods for cross-lingual information retrieval. *In International Conference on Database and Expert Systems Applications*, Springer. pp. 791–801.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3:993–1022.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics.* 19(2):263–311.

Cui, W., S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. 2011. Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions.* 17 (12), 2412–21.

Delpech, E. 2011. Evaluation of terminologies acquired from comparable corpora: an application perspective. *In Proceedings of the 18th International Nordic Conference of Computational Linguistics. (NODALIDA 2011).* pp. 66–73. Riga, Latvia.

Dhillon, P. S., D. P. Foster, and L. H. Ungar. 2015. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research* 16, 3035–78.

Etchegoyhen, T. and A. Azpeitia. 2016. A portable method for parallel and comparable document alignment. *Baltic Journal of Modern Computing* 4 (2), 243.

Fung, P. and P. Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of the 20th international conference on Computational Linguistics (COLING '04).* Stroudsburg, PA, USA, Association for Computational Linguistics. pp. 1051.

Harispe, S., S. Ranwez, S. Janaqi, and J. Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies* 8 (1), 1–254.

Hieber, F. and S. Riezler. 2015. Bag-of-words forced decoding for cross-lingual information retrieval. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado. pp. 1172–1182. Association for Computational Linguistics.

Ion, R., A. Ceauşu, and E. Irimia. 2011. An expectation maximization algorithm for textual unit alignment. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, Oregon. pp. 128–35. Association for Computational Linguistics.

Knoth, P., L. Zilka, and Z. Zdrahal. 2011. Using explicit semantic analysis for cross-lingual

link discovery. *In 5th International Workshop on Cross Lingual Information Access (IJC-NLP 2011)*, Computational Linguistics and the Information Need of Multilingual Societies (CLIA), Chiang Mai, Thailand. pp. 2–10.

Li, B. and E. Gaussier. 2010. 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. Beijing, China. COLING 2010 Organizing Committee.

Li, B. 2012. Measuring and improving comparable corpus quality (Doctoral dissertation. *PhD dissertation*. University of Grenoble. France

Littman, M. L., S. T. Dumais, and T. K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. *In G. Grefenstette (Ed.), Cross-Language Information Retrieval*, Volume 2 of The Springer International Series on Information Retrieval, pp. 51–62. Springer US.

Morin, E., and E. Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, Oregon. pp. 27–34. Association for Computational Linguistics.

Morin, E., Hazem, A., Boudin, F., and Clouet, E. L. 2015. Lina: Identifying comparable documents from wikipedia. *In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, BUCC@ACL/IJCNLP 2015*, pp. 88–91. Beijing, China. Association for Computational Linguistics.

Munteanu, D. S., and Marcu, D. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, vol. 31, no. 4, pp. 477–504. MIT Press.

Oshikiri, T., Fukui, K., and Shimodaira, H. 2016. Cross-lingual word representations via spectral graph embeddings. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, Berlin, Germany, Volume 2: Short Papers. Association for Computational Linguistics.

Otero, P. and López, I. 2011. Measuring comparability of multilingual corpora extracted from Wikipedia. *In Iberian Cross-Language Natural Language Processing Tasks (ICL)*, pp. 8. Published by Paolo Rosso, Alberto Barrón-Cedeño, Marta Vila, Jorge Civera, Anabela Barreiro, Iñaki Alegria.

Pinnis, M., Ion, R., Stefanescu, D., Su, F., Skadina, I., Vasiljevs, A., and Babych, B. 2012 Accurat toolkit for multi-level alignment and information extraction from comparable corpora. *In Proceedings of the ACL 2012 System Demonstrations, ACL '12*, Stroudsburg, PA, USA, pp. 91–6. Association for Computational Linguistics.

Saad, M. 2015. Mining Documents and Sentiments in Cross-lingual Context. *Ph. D. thesis*. Université de Lorraine.

Saad, M., Langlois, D., and Smaïli, K. 2013. Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. *Procedia - Social and Behavioral Sciences* 95 (0), 40–7. Alicante, Spain. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013). Elsevier.

Saad, M., Langlois, D., and Smaïli, K. 2014. Cross-lingual semantic similarity measure for comparable articles. *In Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014*, Warsaw, Poland. pp. 105–15. Springer International Publishing.

Sharoff, S., Zweigenbaum, P., and Rapp, R. 2015. Bucc shared task: Cross-language document similarity. *ACL-IJCNLP 2015*, 74. Association for Computational Linguistics.

Skadina, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. 2012. Collecting and using comparable corpora for statistical machine translation. *In Proceed-*

ings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 438–445 Istanbul, Turkey. European Language Resources Association.

Smith, J., Quirk, C., and Toutanova, K. 2010. Extracting parallel sentences from comparable corpora using document level alignment. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403–11. Association for Computational Linguistics.

Ture, F. 2013. 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.

Vulić, I. and Moens, M.-F. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 349–62. Association for Computational Linguistics (ACL).

Vulić, I. and Moens, M.-F. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15*, New York, NY, USA, pp. 363–72. Association for Computing Machinery.

Wołk, K. and Marasak, K. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, v. 18, pp. 126–132, Elsevier.