HAL
open science

Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities

Motaz Saad, David Langlois, Kamel Smaïli

To cite this version:

Motaz Saad, David Langlois, Kamel Smaïli. Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. Procedia - Social and Behavioral Sciences, 2013, 95, pp.40-47. 10.1016/j.sbspro.2013.10.620 . hal-00907442

HAL Id: hal-00907442

https://inria.hal.science/hal-00907442v1

Submitted on 20 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2013) 000–000

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

V International Conference on Corpus Linguistics (CILC2013)

Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities

Motaz Saad*, David Langlois, Kamel Smaïli

Speech Group, LORIA (UMR7503 : Université de Lorraine, INRIA, CNRS), Villers-lès-Nancy, F-54600, France

Abstract

Parallel corpora are not available for all domains and languages, but statistical methods in multilingual research domains require huge parallel/comparable corpora. Comparable corpora can be used when the parallel is not sufficient or not available for specific domains and languages. In this paper, we propose a method to extract all comparable articles from Wikipedia for multiple languages based on interlanguge links. We also extract comparable articles from Euro News website. We also present two comparability measures (CM) to compute the degree of comparability of multilingual articles. We extracted about 40K and 34K comparable articles from Wikipedia and Euro News respectively in three languages including Arabic, French, and English. Experimental results of comparability measures show that our measure can capture the comparability of multilingual corpora and allow to retrieve articles from different language concerning the same topic.
© 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.

Keywords: computational linguistics ; comparable corpora ; comparability measure

1. Introduction

Parallel and comparable articles are a set of texts in multiple languages, but parallel texts are translation from each others and they are aligned sentence by sentence (aligned at sentence level), while comparable texts are not the exact translation from each others, but they talk about the same topic. So, comparable corpora are a collection of comparable articles aligned at article level.

Parallel or comparable corpora are useful in several domains such as multilingual text mining, bilingual lexicons extraction, cross-lingual information retrieval and machine translation. In these domains, a lot of works use statistical methods. Comparable corpora also can be used to build parallel corpora. Indeed, there are many works about sentences alignment in the comparable corpora (Smith et al., 2010, Do et al., 2010, Abdul Rauf and Schwenk, 2011, Hewavitharana and Vogel, 2011). For these works, comparable corpora must be available. Morevover, statistical methods need huge data, so the more data you have, the better models you can build. Unfortunately, parallel corpora are not available for all domains and languages. They can be extended using human translators, but this is expensive

* Corresponding author. Tel.: +33-383592097; fax: +33-383413079.
E-mail address: motaz.saad@loria.fr

1877-0428 © 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.

[[ar:مطر]]
[[de:Regen]]
[[es:Lluvia]]
[[fr:Pluie]]
[[en:Rain]]

Fig. 1. The form of interlanguage links of Wikipedia

and require a lot of efforts. So, comparable corpora are the best alternative in this case, because they are less expensive and more productive.

Comparable corpora can be obtained easily from multilingual textual contents on the web like newspapers websites, but aligning articles is a challenging task. Wikipedia can be considered as a good and large source for comparable corpora, because it covers many languages and topics.

Our objective is to help a media trackers such as journalists to find web documents related to a same given topic, in several languages. We focus in this work on Arabic, French and English. For that, we need to collect French/English/Arabic comparable corpora, and we need to define comparability measures (CM) between documents.

In fact, there are no available Arabic/other languages pair comparable corpora. Therefore, in Section 2, we propose a method to collect corpora from Wikipedia and Euro News. Collected articles are in Arabic, French, and English languages.

Recent work on comparability measures include (Otero et al., 2011) who proposed a comparability measure for Wikipedia corpus. They considered internal links in articles as the vocabulary that they use to make the comparison. Internal links in Wikipedia's articles are titles for other articles. So, their degree of comparability is defined based on the amount of internal links that can be translated into the target language. In other words, their comparability measure inspects the amount of common internal links between source and target articles. Also, (Li and Gaussier, 2010) defined the degree of comparability for the whole corpus as the expectation of finding, for each source word in the vocabulary of the source corpus, its translation in the vocabulary of the target corpus. They measured the comparability of parallel corpora, Then, they showed how the comparability degree decreased as noisy text added to the parallel corpora.

For our work in this paper, we propose in Section 3, two different comparability measures, which are based on binary and cosine similarity measure. The binary measure requires source/ target texts to be represented as bag of words, while the cosine measure requires source/target text to be represented as vectors. To represent text in vector space model, we use a multilingual document representation model based on wordNet dictionary. We also apply Latent Semantic Indexing (LSI) (Rehurek and Sojka, 2010, Rehurek, 2011). Therefore, to compare documents, unlike (Otero et al., 2011), we take into account the whole contents of the documents. Our work is closer to (Li and Gaussier, 2010) but our methods are not based on a translation table, but on bilingual dictionaries. Moreover, (Li and Gaussier, 2010) work is at corpus level while we propose comparability measure at document level.

The rest of this paper is organized as follows, Section 2 describes our comparable corpora extraction method. Section 3 presents the comparability measures we propose. Section 4 describes the bilingual dictionary. Section Section 5 gives experimental results. Finally, conclusions and future works are stated.

2. Extracting Comparable Corpora

In this section, we present two comparable corpora. The first one is extracted from Wikipedia, and the second one is extracted from Euro News website (`http://www.euronews.com`).

Regarding Wikipedia corpus, we extract it by parsing Wikipedia dumps of December 2011. We extract Arabic, French, and English comparable articles based on interlanguage links. In a given Wikipedia article written in a specific language, "interlanguage links" lead to corresponding articles in other languages. The form of these links is "[[$languagecode : Title$]]". For example, the interlanguage links of the English language article "Rain" are presented in Figure 1. We name these corpora as AFEWC.

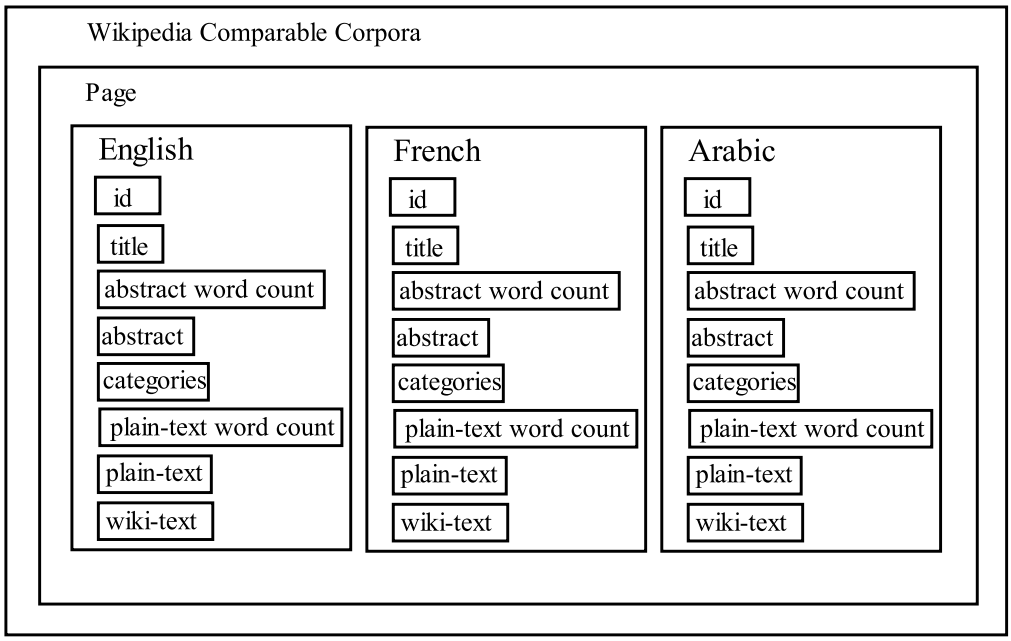

Fig. 2. The Extracted Wikipedia Comparable Corpora AFEWC

Extracted articles are aligned at article level. The extracted information include article's title, and wiki markup. From wiki markup, we extract the article's summary (abstract), categories, and the plain text. We structure all extracted information into XML files as shown in Figure 2. The reason of keeping wiki markup is to extract additional useful information later such as info boxes, image captions. The extraction steps are described below.

For each English article in Wikipedia dump do the following:

1. If the article contains Arabic and French interlanguage links, then extract the titles from interlanguage links for the three comparable articles.
2. Search by titles for the three comparable articles in the Wikipedia dump, and then extract them.
3. Extract the plain-text from wiki markup of the three comparable articles.
4. Write comparable articles into plain-texts and xml files.

Wikipedia December 2011 dumps contain 4M English articles, 1M French articles, and 200K Arabic articles. The extracted comparable articles in AFWEC include 40,290 aligned articles.

Regarding Euro News corpus, we extract it by parsing the html of each English news article, which contains hyperlinks to Arabic and French news articles. Then, we strip html tags for the three comparable articles, and write them into plain text files. We collected 34,442 comparable articles, we name this corpus as eNews. Same as AFEWC, eNews is aligned at article level.

The number of articles, sentences, average sentences per article, average words per article, words, and vocabulary of AFEWC and eNews are presented in Table 1. Both AFEWC and eNews are available online at `http://crlcl.sf.net`. These resources are interesting for many applications such as statistical machine translation, bilingual lexicons extraction, cross-language information retrieval, and comparing multilingual corpora. AFEWC is also interesting for text summarization because article's summaries (abstracts) are written by Wikipedia contributors which make them high quality summaries, they can be also used as benchmarks for text summarization researches.

Table 1. AFEWC and eNews corpora information

	AFEWC			eNews		
	English	**French**	**Arabic**	**English**	**French**	**Arabic**
Articles	40290	40290	40290	34442	34442	34442
Sentences	4.8M	2.7M	1.2M	744K	746K	622K
Average #sentences/article	119	69	30	21	21	17
Average #words/article	2266	1435	548	198	200	161
Words	91.3M	57.8M	22M	6.8M	6.9M	5.5M
Vocabulary	2.8M	1.9M	1.5M	232K	256K	373K

3. Comparability Measures

As we stated in the introduction, comparable articles are nearly equivalent text in different languages. But we do not know the comparability degree of these articles. So, we propose in this section two comparability measures (CM) for comparable articles. The CM ranges from 1 (fully parallel) to 0 (not parallel nor comparable). When the measure is near 0, then the articles talk about different topics.

Binary and cosine measures are common methods for measuring similarity. In our work, we measure the comparability of articles using these measures. Binary comparability measure (binCM) requires source/target text to be represented as bag of words (BOW), while cosine comparability measure (cosineCM) requires source/target text to be represented as vectors. A bilingual dictionary can be used to align source/target words of comparable articles, but texts must be adapted to dictionaries in order to obtain good coverage (lemmatization,...). In the following, we start first by defining our measures and multilingual document representation models, then, we describe the bilingual dictionary that we use, and the applied morphological analysis for words.

Regarding binCM, we first define the binary function $trans(w_s, d_t)$ that returns 1 if a translation is found in a bilingual dictionary for the source word w_s in the target document d_t, and 0 otherwise. So, binCM for source and target articles, d_s and d_t, is defined as follows:

$$binCM(d_s, d_t) = \frac{\sum\limits_{w_s \in d_s \cap V_s} trans(w_s, d_t)}{|d_s \cap V_s|} \tag{1}$$

where V_s is the source vocabulary of the bilingual dictionary, d_s and d_t are the source and target documents, considered as bags of words. Because $binCM(d_s, d_t)$ is not symmetric, the actual value used for measuring the comparability between d_s and d_t is as follows:

$$\frac{binCM(d_s, d_t) + binCM(d_t, d_s)}{2} \tag{2}$$

Regarding cosineCM, we need to represent source/target texts as vectors. In our work, we represent multilingual document in the Vector Space Model (VSM). This is done by using a bilingual dictionary to align words in source/target vectors. Additionally, Latent Semantic Indexing (LSI) (Rehurek and Sojka, 2010, Rehurek, 2011) can be used to reduce the dimensionality of VSM. We start first by defining cosine comparability measure, then we describe the VSM in details later.

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. Cosine similarity is often used to compare documents in text mining. Given two vectors A and B of attributes representing the source and target documents, the cosine similarity $cosineCM(A, B)$ between A and B is represented using a dot product and magnitude:

$$cosineCM(A, B) = \frac{A \cdot B}{||A|| \times ||B||} = \frac{\sum\limits_{i=1}^{n} A_i \times B_i}{\sum\limits_{i=1}^{n} (A_i)^2 \times \sum\limits_{i=1}^{n} (B_i)^2} \tag{3}$$

The resulting similarity ranges from -1 which means exactly opposite, to 1 which means exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity. For text matching, the attribute vectors A and B are usually the weights of words in the comparable articles. Since the weights of words are positive values, then the cosine of two documents ranges from 0 to 1: that matches with our definition of a comparability measure.

To represent multilingual articles in the vector space model (VSM), we build the source and target vectors with the following method: we use a bilingual dictionary; for each translation $a \rightarrow b$ in this dictionary, the vectors have one attribute. For the source vector this attribute is equal to the frequency of a in the source document (0 if a is not in the document), and for the target vector this attribute is equal to the frequency of b in the target document (0 if b is not in the target document). We name this representation as VSM-DIC.

Latent Semantic Indexing (LSI) can be applied to build VSM and to make mapping into a new reduced space. We start with a matrix. There is one column for each source and one column for each target word in the corpora. There is one line for each source document, and one line for each target document. The intersection between a column and a line contains the frequency of the corresponding word in the corresponding document. Note that this matrix is sparse. Then we apply LSI to map this matrix in a more compact vector space model. It is then possible to compute the vector corresponding to a document into this vector space model and apply formula 3. We name this representation as VSM-LSI.

4. Dictionary and coverage

Regarding the bilingual dictionary that we use for word alignment, we use Open Multilingual WordNet (OMWN) (Bond and Paik, 2012) which is available in many languages. From OMWN, we extracted 148,730 English words and 14,595 Arabic words. All these words are listed in different sets of synonyms (named 'synsets'). These sets are very useful because we use them to look for possible translations from source to target.

Because the bilingual dictionary does not cover all word variations and morphologies, we apply morphological analysis to words in documents in order to increase the coverage of dictionary between source and target texts. We first remove stopwords from all comparable articles before processing them.

There are many word reduction techniques which are language dependant. For English, stemming and lemmatization are widely used and known techniques in the community. Stemming (Porter, 2001) truncates a word into a stem, which is a part of the word, and may not be in the dictionary, while lemmatization (Stark and Riesenfeld, 1998) retrieves the dictionary form (lemma) of an inflected word.

Table 2. Morphology richness for Arabic language

Arabic word	English meaning	Description
كتب *ktb*	to write	the root
كَاتب *kātb*	an author	a name of the subject
يكتب *yktb*	he writes	the verb
كتَاب *ktāb*	a book	the outcome of the verb
مكتبة *mktbh*	library	the place of the verb (to put the outcome)
مكتب *mktb*	office	the place of the verb (to write)
طير *ṭyr*	to fly	the root
يطير *yṭyr*	he flies	the verb
طَائِر *ṭā'ʾr*	a bird	a name of the subject
طيَار *ṭyār*	a pilot	a name of the subject
طَائِرة *ṭā'ʾrh*	an air plane	a name of the subject

Arabic morphology is rich because Arabic words have a root and prefixes, infixes, and suffixes and we can inflect many words from this root. For Arabic, rooting and light stemming (Saad and Ashour, 2010) are widely used. In the

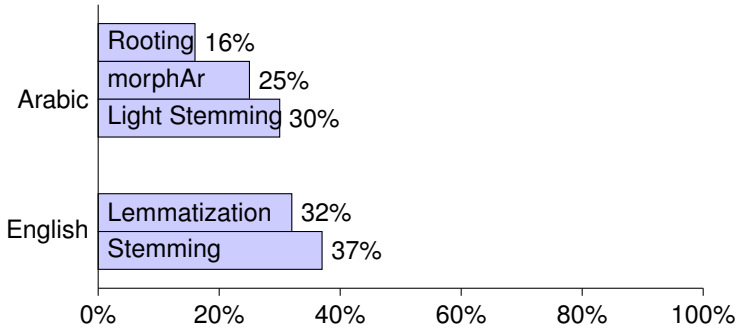

Fig. 3. OOV rate of Word reduction techniques for Arabic and English for parallel corpora

following, we describe rooting and light stemming for Arabic language. Rooting removes word's affixes, then convert to root, while light stemming just removes word's affixes. In Arabic language, word variants usually do not have similar meanings or semantics although these word variants are generated from the same root. Thus, root extraction algorithms affect the meanings of words if we consider their meanings in other languages. Arabic light stemming aims to retain the words meanings. To see how Arabic light stemming retains words meaning, consider the following Arabic word examples, the words (المكتبة *ālmktbh* "the library"), (الكَاتب *ālkātb* "the author"), (الكتَاب *ālktāb* "the book") belong to the same root (كتب *ktb* "to write"), in spite they have different meanings in English language. Arabic rooting reduces their semantics by converting them to their root. Arabic light stemming, on the other hand, maps the word (المكتبة *ālmktbh* "the library") to (مكتبة *mktbh* "library"), and the word (الكتَاب *ālktāb* "the book") to (كتَاب *ktāb* "book"), and the word (الكَاتب *ālkātb* "the author") to (كَاتب *kātb* "author"). Another example is the Arabic root word (سفر *sfr* "to travel") which can be inflected to two plural forms (المسَافرون *ālmsāfrwn*) in nominative form, and (المسَافرين *ālmsāfryn*) in accusative/genitive form, which both mean "the travelers". Arabic light stemmer reduces them both to (مسَافر *msāfr* "traveler"). Also, the word (سيسَافر *sysāfr* "he will travel"), its light stem is (سَافر *sāfr* "travel"). The reader can find these examples and another ones in Table 2 which shows the richness of Arabic morphology and how rooting can change the corresponding meanings in other languages.

In order to increase the dictionary coverage of such rich morphology while not losing words meaning, we have developed a reduction technique for Arabic words, which combines rooting and light stemming techniques. We name this technique as morphAr. The idea of morphAr is to try to reduce a word using light stemming first. If we obtain a reduced form, present in the dictionary we stop, and if not, we apply rooting to reduce the word.

In Figure 3, we measure the Out of Vocabulary (OOV) rate for all the presented word reduction techniques. If we consider word reduction techniques for each language separately, then rooting for Arabic and lemmatization for English have the lowest OOV rate. But we do not aim to just reduce OOV independently for each language, we aim to increase alignment rate for source/target words by finding the appropriate translation for these words in the bilingual dictionary. We experienced different combinations of word reduction techniques in both Arabic and English as presented in Figure 4, and we find that morphAr for Arabic and lemmatization for English lead to the best coverage in the bilingual dictionary. So, we use this combination of techniques in the following experiments.

5. Experimental results

We applied our comparability measures on parallel and comparable corpus. For comparable corpus, we merged AFEWC and eNews. For parallel corpus, we merged several corpora: AFB and ANN which are provided by LDC `http://ldc.upenn.edu`.

First we show that our comparability measures give better score to parallel corpora than to comparable corpora. This is a way to check that our measures work well. Indeed, articles in parallel corpus tend to be more close in terms of contents than comparable corpus, because parallel corpora are aligned at sentence level.

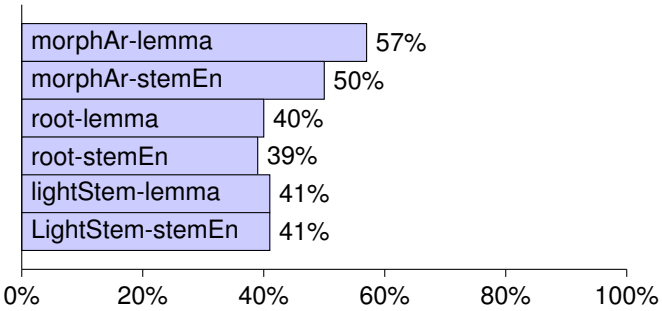

Fig. 4. Coverage rate of the bilingual dictionary for combined English/Arabic morphological analysis techniques

Table 3 presents the results of binCM, and cosineCM using VSM-DIC and VSM-LSI. For each comparability measure, we compute the average score of comparability across article pairs in the corresponding corpus. First, results confirm that CM for parallel corpus is larger than the one for comparable corpora. Second, results also show that cosineCM based is better than binCM. This was expected because cosineCM uses richer information from the documents than binCM. Last, VSM-DIC and VSM-LSI obtained the same results, this is due to presentation of English/Arabic articles to build VSM matrix, that is, presenting English and Arabic articles separately in the matrix makes LSI not able to find the semantic relation between Arabic/English words.

Table 3. CM computed for parallel and comparable corpora

		cosineCM	
	binCM	**VSM-DIC**	**VSM-LSI**
Parallel	0.21	0.26	0.26
Comparable	0.11	0.18	0.18

However, this evaluation deals only with average scores. We recall that the objective is to retrieve target from comparable articles from a given source article. So, to measure this capability of our CM, we compute the recall. Multilingual information recall is computed by providing each source article as query. For this source article, we evaluate all the target articles with the CM. Then we select the best target articles according to their scores. Then, we check if the correspondent target article is in the 1-top list (first recall R1), the 5-top list (fifth recall R5), and the 10-top list (tenth recall R10). To perform this evaluation, we select 100 source (English) articles from comparable corpora to retrieve corresponding target (Arabic) comparable articles.

Figure 5 presents the recall for each size of top-list (percentage of times the correspondent Arabic article is in the top list) for our measures. The figure shows that cosineCM achieves the highest performance, while binCM achieves the lowest performance. However, R10 is perfect for all measures, this is due to the fact that we make retrieval among 100 articles only. This confirms the advantage of VSM for multilingual articles. We also note here that VSM-DIC and VSM-LSI has the same performance due to the reasons described earlier.

6. Conclusions

In this paper, we introduced comparable corpora extracted from Wikipedia and Euro News in Arabic, French, and English which can be considered as an interesting linguistic resources for statistical machine translation, text summarization, multilingual text mining, information retrieval lexicons extraction. Besides extracting the comparable corpora, we also proposed two comparability measures for the comparable corpora which measure the degree of comparability of multilingual articles. Experimental results showed that our proposed measures are reliable and can capture the comparability degree for our comparable corpora. In the future work, we will improve multilingual documents representation for enable LSI to find semantic relations of words from different languages. we will also

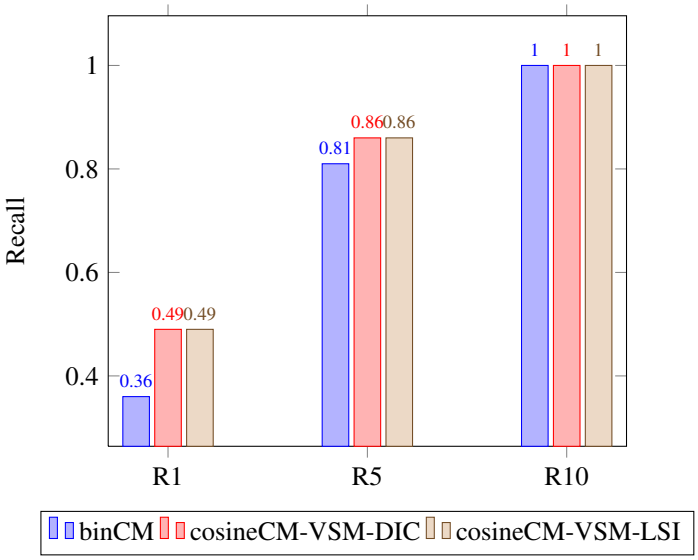

Fig. 5. CM evaluation on Comparable corpora

investigate statistical machine translation approach to measure comparability. In addition, we will discover opinions across comparable articles in our corpora.

References

Abdul Rauf, S., Schwenk, H., (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine translatio*n, 1–35.

Bond, F., Paik, K., (2012). A survey of wordnets and their licenses. In: *6th Global WordNet Conference (GWC2012)*. p. 6471.

Do, T. N. D., Besacier, L., Castelli, E., (2010). A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora. In: *European Conference on Machine Translation (EAMT) 2010*. Saint-Raphael (France).

Hewavitharana, S., Vogel, S., (2011). Extracting parallel phrases from comparable data. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. BUCC '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 61–68.

Li, B., Gaussier, E., (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 644–652.

Otero, P., López, I., Cilenis, S., de Compostela, S., (2011). Measuring comparability of multilingual corpora extracted from wikipedia. In: *Iberian Cross-Language Natural Language Processings Tasks (ICL)*. p. 8.

Porter, M., (2001). Snowball: A language for stemming algorithms.

Rehurek, R., (2011). Subspace tracking for latent semantic analysis. In: *Proceedings of the 33rd European conference on Advances in information retrieval*. ECIR'11. Springer-Verlag, Berlin, Heidelberg, pp. 289–300.

Rehurek, R., Sojka, P., (2010). Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.

Saad, M., Ashour, W., (2010). Arabic morphological tools for text mining. In: *EEECS10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*. European University of Lefke, Cyprus, pp. 112–117.

Smith, J., Quirk, C., 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. Association for Computational Linguistics*, pp. 403–411.

Stark, M., Riesenfeld, R., (1998). Wordnet: An electronic lexical database. In: *Proceedings of 11th Eurographics Workshop on Rendering*.