Measuring the comparability of multilingual corpora extracted from Twitter and others

Abidi Karima, Kamel Smaili

Measuring the comparability of multilingual corpora extracted from Twitter and others

ABIDI Karima[1] and Kamel Smaïli[2]

[1] Ecole Suprieure d'Informatique (ESI) Algiers,Algeria
[2] Campus Scientifique LORIA, Nancy France

Abstract. Multilingual corpora are widely exploited in several tasks of natural language processing, these corpora are principally of two sorts: comparable and parallel corpora. The comparable corpora gather texts in several languages dealing with analogous subjects but are not translations of each other such as in parallel corpora. In this paper, a comparative study on two stemming techniques is conducted in order to improve the comparability measure based on a bilingual dictionary. These methods are: Buckwalter Arabic Morphological Analyzer (BAMA) and a proposed approach based on Light Stemming (LS) adapted specifically to Twitter, then we combined them. We evaluated and compared these techniques on three different (English -Arabic) corpora: a corpus extracted from the social network Twitter, Euronews and a parallel corpus extracted from newspapers (ANN). The experimental results show that the best comparability measure is achieved for the combination of BAMA with LS which leads to a similarity of 61% for Twitter, 52% for Euronews and 65% for ANN. For a confidence of 40% we aligned 73.8% of Arabic and English tweets.

Keywords: Twitter, Stemming, comparability measure

1 Introduction

Several applications in natural language processing necessitate parallel corpora. For classical issue of machine translation, this type of data is available: Europarl [9], Hansard[13] and Hong Kong [16]. Unfortunately, this kind of data is not available for all languages and especially for under-resourced ones and those used in social networks where the users allow themselves a certain freedom of writing. That is why nowadays, researchers make efforts to investigate comparable corpora because they are more available than parallel corpora which need a human effort to align corresponding sentences.
Comparable corpora require alignment procedure to retrieve parallel segments within these documents. Consequently, we should determine which document is comparable to another one. To achieve this previous task, we need first to define precisely the concept of comparability.
Li [10] defines the comparability as follows: *Two corpora in two different languages L_1 and L_2 are considered as comparable, if the translation of a part of L_1 respectively L_2 does exist in L_2 respectively L_1*. Comparable corpora can be collected from different Internet sources such as newspapers, Wikipedia dump files and from Social networks, but the alignment of these documents remains a challenge in natural language processing.
Nowadays, Twitter has became a popular social network which allows millions of users to share their opinions on various topics. The emergence of this kind of data has led to many research works with the objective to exploit this large quantity of information to build monolingual and multilingual corpora [6], [12]. Another research activity that exploits social networks concerns multilingual sentiment analysis and sentiment classification from tweets [2], [5], [4].

Our objective in this paper is to collect comparable data about Syrian war and refugee problems by crawling a social network in Arabic and English. Obviously, we can not just crawl Twitter, we need to compare the retrieved corpora in order to build a real comparable corpus. For that, we have to measure the comparability between the English and the Arabic corpora. The most used methods for this issue are based on a bilingual dictionary. The drawback of this method is that the bilingual dictionaries can not cover the whole existing corpora and especially those extracted

from social networks. Consequently, the dictionary should cover largely the processed corpora. Otherwise, the comparability measure will be biased by the weak rate of coverage of the bilingual dictionary.
To increase the coverage of the bilingual dictionary, we will use stemming methods in order to rewrite the corpora in terms of lemmas especially for Arabic. For this language, we will be faced, in addition to the classical problems of processing Modern Standard Arabic, to the noisy text which includes misspelled words, Arabic dialects phrases, the stretched or duplicated letters, etc.

This paper is organized as follows: in section 2 we summarize some previous work on comparability measures. We describe the collected multilingual corpora and the preprocessing steps applied to each corpus in Section 3. In section 4 we described a stemming method used by the community (BAMA), we propose a light stemming method and then we combine the two previous approaches. In section 5 we present Li and Gaussier comparability measure that we need for retrieving a comparable documents. Several experiments and results are described in section 6. Finally, we conclude and present some future works.

2 Previous works

The construction of comparable corpora is performed using similarity measures. These measures can be based on three different approaches: vocabulary overlapping, vectors space and Machine translation. Among existing work to align the comparable corpus we can cite the following: Li and Gaussier in [11] defined the degree of comparability between two corpora as the expectation of finding, for each word of the source corpus, its translation in the target one. They use this definition to propose a measure which estimates the comparability of a parallel corpus, then they showed how the comparability degree has decreased when noises have been added to the parallel corpora. Then this measure has been modified, to take into account the context in order to reduce the issue of polysemy.
Li and Gaussier's measure has been improved by Guiyao et al in[8] by taking into account the occurrence of a word concerned by the comparability and its different translations.
Saad et al [14] used LSA for measuring the comparability between two corpora. They proved that this approach gives better results than those based on the common vocabulary.
Gamallo in [7] proposed two strategies to build comparable corpora from Wikipedia, they used (binary and TFIDF) Dice to measure the degree of comparability of 30 different comparable corpora. They proved that the Dice score based on TFIDF achieves better results than those based on the binary representation.

3 Corpora and preprocessing

To study the reliability of the techniques proposed in this work, we use several corpora extracted from: Twitter, Euronews and a parallel corpus (ANN)[15].
Our objective is to retrieve comparable corpora from Twitter, this could be considered as a difficult task. In this work, we validate the proposed approach on supposed easier corpora Euronews (comparable) and ANN (parallel).

3.1 Twitter corpus

The objective is to build an English-Arabic comparable corpus. For that, we extract from Twitter by using Talend Studio and Twitter API the most frequent Hashtags for each language concerning the topics *War in Syria* and *The Syrian Refugees*. Table 1 and Table 2 show these Hashtags and the corresponding number of the retrieved tweets. We also used related Hashtags to enlarge our corpus, for example: #Syrianarmy, #freesyrianarmy. Due to the particularity of Arabic, we used different Hashtags such as: #سوريا, #سورية, both of them correspond to the word *Syria*. In fact, this word could be written in two ways in MSA (Modern Standard Arabic) with symbol *Ta* (ة) and with symbol *alif* (ا).

Table 1. Number of English tweets collected for each Hashtag.

English Hashtag	N_{Twts}
#SyrianRefugees	10895
#refugeescrisis	2856
#Syrianarmy #freesyrianarmy	3760
#SyriaCrisis	6260
#syria #syrian	34000
Total	**57771**

Table 2. Number of Arabic tweets collected for each Hashtag.

Arabic Hashtag	Translation	textbfN_{Twts}
#اطفال_سوريا	Children of Syria	1599
#الثورة_السورية	Syrian revolution	10092
#اللاجيئين_السورين	Syrian refugees	4000
#سوريا, #سوريه	Syria	34000
#الجيش_السوري_الحر, #الجيش_العربي_السوري	Syrian arab army, Free Syrian army	5234
Total		**59452**

Table 3 shows the characteristics of all corpora used in this work, where $|S|$ is the number of sentences, and $|V|$ is the vocabulary size.

Table 3. Comparable and parallel corpora characteristics.

	Twitter		**Euronews**		**Parallel**	
	English	Arabic	English	Arabic	English	Arabic
$\|S\|$	12624	14412	1400	1400	300	300
$\|V\|$	108K	372K	145K	378K	119K	377K

3.2 Preprocessing corpora

Natural Language preprocessing for MSA (Modern Standard Arabic) are available, but the challenge in this work is the nature of the processed data. In fact, tweets are short messages, restricted to 140 characters, these messages often contain misspelled words, abbreviations and dialectal words where the analyzer is unable to produce the morphological form of these messages. Furthermore, Arabic language is morphologically rich because in the majority of cases, a word is composed by the concatenation of a root and affixes. A root in Arabic is considered as a producer of words, that is why from a single root, several words can be composed. For example: the root كتب *to write* with specific affixes produces different words with different meanings: يكتب *he writes*, مكتبة *library*, مكتب *office*,..etc.

In the following we will present the different cleaning operations related to English and Arabic.

3.2.1 Preprocessing English We present below several procedures applied to the English part of Twitter corpora.

- Since Twitter is used by million of users, people take some freedom in writing. That is why we use an SMS dictionary [3] which contains abbreviations, acronyms and their literal corresponding text. We use this dictionary to replace abbreviations by their literal forms. For example *ppl* will be replaced by *people*.
- In order to remove foreign Twitter in the English part of the corpora, we developed a tool based on stop words and ASCII code to filter the foreign phrases.
- Contrary to Arabic corpus, the English one contains many Hashtags embedded in the Tweet itself. A Hashtag is composed of several words *#SyrianArabArmy, #SyrianCivilians, ...etc* which should be split in order to obtain a larger Tweeter vocabulary. In the case where words are separated by capital letters or special characters it is easy to determine the words composing the Hashtag, such as in the examples: *#SyrianArabicArmy #Syrian_Arabic_army*. But, when the Hashtag is completely written in lowercase, we use a dictionary of words, sorted alphabetically and by length which allows to find the different words which match best with the content of the Hashtags.

3.2.2 Preprocessing Arabic For Arabic, the preprocessing is achieved in order to reduce the ambiguity due to the fact that people write some words differently. In the following we present some rules used for that issue.

- We replace the letter ة by ه only at the end of words. This could be very surprising since these two letters are grammatically different, but graphically they are similar except that the first one has two dots above. For convenience, some people use indifferently both of them. This is why we homogenize the script in all the tweets. For almost the same reason, we replace all the forms of symbol *Alif* with *Hamza* such as in: آ, أ, إ with a simple *Alif* ا.
- All the diacritics have been removed since people can read without these vowels.
- In social networks, sometimes users stretch words to accentuate their opinion or just write the words such as they pronounce them, we need to normalize the way of writing these words. The stretched or duplicated letters are removed such as in: عاااجل → عاجل and يااارب → يارب.

In opposite to Twitter corpora, Arabic parallel and comparable corpora used in this work do not need all these preprocessing steps, because these corpora are collected from newspapers. Consequently, they do not contain noisy data as in Twitter corpora. Nevertheless, we removed special characters, Latin characters and stops words. We also transform as above all the *Alif* with *Hamza* by a simple *Alif*.

4 Stemming

For documents such as Twitter, the vocabulary is not large, in order to use statistical methods, in general we segment the words by using lemmatization or stemming. The idea is to replace different words which share the same root by this root itself. These methods consist in reducing inflectional and sometimes derivational forms of a word to a common base form. In Arabic, this task becomes more complex compared to other languages, because, this language has some particularities such as agglutination, using diacritics, etc. In fact, Arabic words can be formed by attaching affixes to a root. Affixes in Arabic are: prefixes, antefixes, suffixes and postfixes. Prefixes and antefixes are attached at the beginning of the words, where suffixes and postfixes are attached at the end. For example the word ليفاوضونهم which means *to negotiate with them* is composed by the elements shown in Table 4. To improve the dictionary coverage rate, in this work, we applied different techniques to retrieve the most representative form of words. For English corpora, we used TreeTagger [4] and for Arabic corpora, we used the two following methods.

[3] http://www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm
[4] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Table 4. Example of Arabic Affixes

Word	Antefixe	Prefixe	Stem	Suffixe	Postfixe
ليفاوضونهم	ل	ي	فاوض	ون	هم

4.1 Buckwalter Arabic Morphological Analyzer

This analyzer is one of the most referenced in the literature, coded with Perl language by Tim Buckwalter [3]. It is designed as a main database of 40,648 lemmas supplemented by three morphological compatibility tables used for controlling affix-stem combinations [1]. For each word entered, BAMA provides: the different possible lemmas of the word with its diacritics, grammatical label of each lemma and their corresponding translation in English, as shown in Figure 1.

Fig. 1. BAMA output of Arabic word يتكلمون *they speak*

<morphology_analysis total_words="1" >
<word total_analysis="1" value="يتكلمون" w_id="1">
<analysis additional_info=Lemma:تَكَلَّمَ Gloss: prefix : they (people) + stem : speak/talk/discuss + suffix : [masc.pl.] suffix=[وُنَ :IV SUFF_SUBJ:MP_MOOD:I] prefix="[يَ :IV3MP] type="" impartial="" transitive="" pos=VERB_IMPERFECT root="" pattern="" stem=تَكَلَّمَ vowled= يَتَكَلَّمُوُنَ a_id="1"/>
</word>
</morphology_analysis >

4.2 Method based on Light Stemming (LS)

Arabic Light stemming consists in removing all affixes attached to the Arabic word. However, this approach may face to several ambiguities. For instance, when a word contains a sequence of letters that seems to be an affixe but which is not. For example, the sequence ان in the word: طالبان *two students* indicates the dual-representation of the word student. However, the same sequence of letters in the word سكان *residents* is not an affix but it is a part of this word, its deletion will produce a false lemma.
That is why, we proposed a new method of light stemming, which removes all affixes attached to the Arabic word according to a specific order. We start by removing the external affixes of the word then those connected directly to the stem, as shown in Figure 5. Each time an affix is removed the remaining lemma is kept for the further processing steps as described in Algorithm 1 and 2. Using the affix categories shown in table 5, Algorithm 1 removes firstly, only the Antefix (

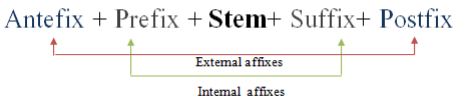


Fig. 2. Topology of an Arabic word

RAntefix(W)) attached to the Arabic word, secondly only the Postfix (RPostefix(W)) and finally, it removes both of them (RPostfixAntefix(W)). After each removal operation, the achieved lemma is kept. The same procedure is then applied to remove the prefix (RPrefix(l_i)), suffix (RSuffix(l_i)) and finally, it removes both of them (RPrefixSuffix(l_i)) from each lemma kept in the previous step.

Algorithm 1 Generation of lemmas

$S_l \leftarrow \emptyset$
//**Removing External affixes** by using $T_{Antefix}$ and $T_{Postfix}$ affixes tables:
$S_l \leftarrow$ RAntefix(W);
$S_l \leftarrow$ RPostfix(W);
$S_l \leftarrow$ RPostfixAntefix(W);
//**Removing Internal affix** by using T_{Prefix} and T_{Suffix} affixes tables:
for each l_i in S_l **do** // $i = 1 \dots 3$
$S_l \leftarrow$ RPrefix(l_i);
$S_l \leftarrow$ RSuffix(l_i);
$S_l \leftarrow$ RPrefixSuffix(l_i);
end for
Filtering (S_l)

Consequently, for each word, we get twelve lemmas, they may contain wrong and duplicated segmentation hypotheses. The filtering algorithm will remove the duplicated lemmas and those which are not present in a huge corpus [5] as explained in Algorithm 2.

Algorithm 2 Filtering(S_l)

$S_{Out} \leftarrow \emptyset$
$S_l^{'} \leftarrow RemoveDuplicatedLemma(S_l)$
for each l_i in $S_l^{'}$ **do** $F \leftarrow CalculateFrequency(l_i, C_A)$ // C_A is a large corpus
if F ¿ 0 **then** $S_{Out} \leftarrow l_i$
end if
end for

Table 5. Affixes categories most used in Arabic.

	Arabic affixes
$T_{Antefix}$	ل , ب , و ,ف , ك ,وس ,فل ,فب , فس ,لّل ,ول ,و ب ,ال , ولّل , كال فال , وال , وبال
T_{Prefix}	ا, ن, ي, ت
T_{Suffix}	ت, ن, ا, ي, و,ين, وا, تا, تم, تن, نا, ات, ان, ون, تما, يون, تين, تان
$T_{Postfix}$	ي, ه, ك, كم, هم, نا, ها, تي, هن, كن, هما, كما

4.3 Combining BAMA and LS

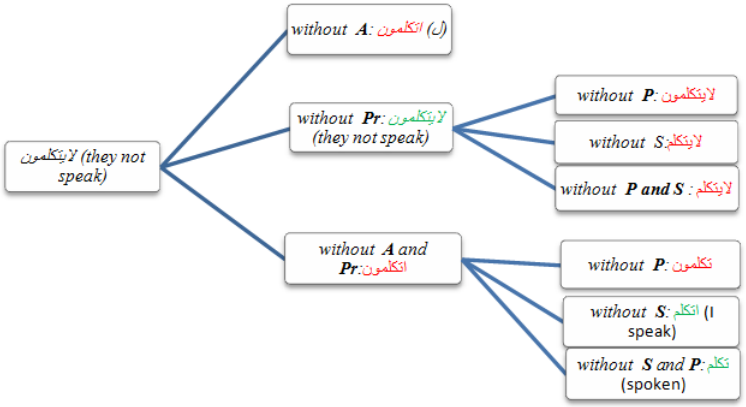
Because BAMA processes a standard Arabic, consequently it fails in analyzing misspelled words which are very frequent especially in Twitter. In the following we give some examples of them:

[5] Corpora: Mourad Abbas (14M) and Saad Motaz(15M)http://aracorpus.e3rab.com

- لايتكلمون *they do not speak* is actually constituted with two words لا followed by يتكلمون which should be separated by a space. This is not the case that makes BAMA failing in the analysis.
- The word باعتبرها should be written such as باعتبارها. Obviously, BAMA is unable to analyze this misspelled word.
- Other words in Twitter are written in Syrian dialect such as : بتقصف *bombarding*. This is not supported in BAMA.
- Some forms of imperative verbs such as: اقصفوا *bomb* are not used in BAMA [1].

In Figure 3, we present how LS can analyze the word لايتكلمون for which BAMA fails. The words

Fig. 3. Lemmas produced by LS for incorrect word لايتكلمون *They do not speak*, A: Antefix, Pr: Prefix, S: Suffix, P: Postfix.



in red indicate wrong lemmas and in green correct lemmas. Contrary to BAMA, LS gives for this word two lemmas: اتكلم *I speak* and تكلم *He speaks*.

LS also has some limitations, for instance, it is unable to analyze correctly the agglutinated words including plural forms. It is able to segment them by identifying the pronoun and the plural form but it fails to identify the root form such as in the example: قلوبهم *their heart*, where LS provides only قلوب and هم.
To overcome the drawbacks of the two methods, we propose to combine them. The limitations of both methods justifies the use of a hybrid method based on BAMA and LS. This justification is strengthened by the fact that BAMA is able to analyze correctly 88% of the content of our Twitter Arabic corpus while LS processes correctly 91%.

In this hybrid method, both approaches cooperate in order to produce all the potential lemmas for each word in order to increase the coverage of the dictionary.

5 Estimating the comparability

In this work, we used Li and Gaussier measure [11] to study the impact of the methods of stemming on the comparability measure. This measure defined the similitude between two corpora as follows: Let assume that C is an Arabic-English corpora consisting of an Arabic part C_a and an English part C_e. The comparability measure can be defined as the expectation of finding, for each English word W_e(respectively W_a) in the vocabulary V_e of C_e (respectively V_a of C_a), its translation in the vocabulary V_a of C_a (respectively V_e of C_e).

The comparability measure is estimated as follows:

$$LG = \frac{\sum_{w \in V_e \cap D_e} \sigma(w, V_a) + \sum_{w \in V_a \cap D_a} \sigma(w, V_e)}{\mid V_e \cap D_e \mid + \mid V_a \cap D_a \mid} \quad (1)$$

where D_e (respectively D_a) is the English side (respectively Arabic side) of a bilingual dictionary.

Let σ a function which indicates whether a translation of a word w represented by a list of potential translations $T(w)$ does exist in the vocabulary V_x.

$$\sigma(w, V_x) = \begin{cases} 1 \ if \ T(w) \cap V_x \neq \emptyset \\ 0 \ else \end{cases} \quad (2)$$

5.1 The coverage rate

We study hereinafter the influence of stemming techniques on the dictionary coverage rate, this parameter is important which influences directly the comparability of corpora as it will be shown in the following experiments. We define the coverage rate of a dictionary D compared to a corpus C represented by a vocabulary V by the quantity:

$$C_v(V, D) = \frac{|V \cap D|}{|V|} \quad (3)$$

Because, we work in a bilingual context (English-Arabic), we calculated a symmetric coverage rate $C_{v_{sy}}$ as follows:

$$C_{v_{sy}}(V, D) = \frac{C_v(V_s, D_s) + C_v(V_t, D_t)}{2} \quad (4)$$

Where: v_s (respectively v_t) the source and the target vocabulary and D_s (respectively D_t) the source and the target part of the bilingual dictionary.
For all experimentation, we calculated the dictionary coverage rate by using OMWN (Open Multilingual WordNet) which contains 17 785 pairs of Arabic and English entries.

6 Experimental Results

In the following we describe the experiments we conducted. In the first one, no stemming has been used to process the English part of Twitter while for the Arabic part, we used BAMA and LS in order to study the impact of these stemming techniques on the coverage rate.
In the second one, we used comparable and parallel corpora collected from newspapers to conduct similar experiments except that in this case we have more clean corpora than Twitter.
For both experiments we calculated the comparability measure between English and Arabic for both categories of corpora.
Figure 4 illustrates the impact of using stemming techniques on the coverage rate of the bilingual dictionary on the Arabic Twitter corpus. Obviously, for Arabic and more especially for Twitter, this result is expectable, in fact the stemming procedure is very useful and necessary.

In Figure 5, we introduce the stemming process (by using TreeTagger) on the English part of Twitter. We show that the stemming for English is also important since the results of coverage have been improved respectively for BAMA with 2.6% and for LS with 2.9%.

Since, the coverage rate is important in the calculation of the comparability measure, we decided to estimate the coverage for corpora which are supposed to be more clean than Twitter

Fig. 4. The dictionary coverage rate by using BAMA and LS stemming techniques only on Arabic part of Twitter corpora.

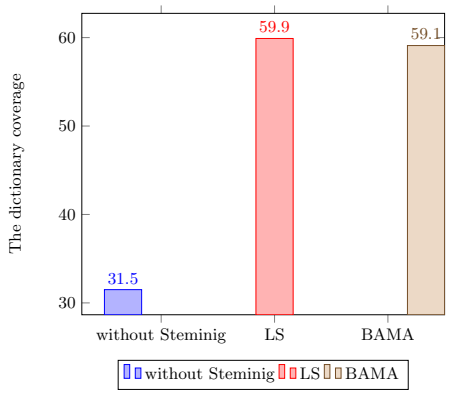


Fig. 5. The dictionary coverage rate by using BAMA and LS stemming techniques on Arabic part and Treetagger for English part of Twitter

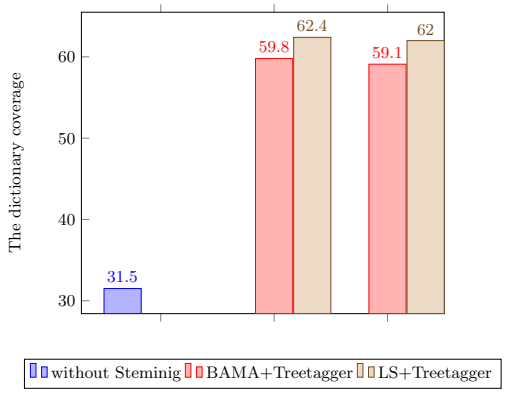
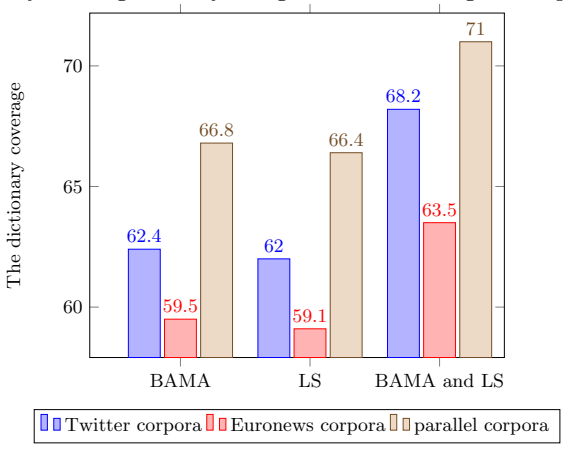


Fig. 6. The dictionary coverage rate by using all the stemming techniques on all corpora.



and for which the documents are strongly related. In Figure 6, we illustrate the coverage rate for three different corpora, the previous Twitter corpus, a comparable one (Euronews) and a parallel corpus (ANN).

This figure shows that the combination of BAMA, LS and TreeTagger achieves the best results whatever the corpus used. The second conclusion is that the best coverage rate is obtained for the

parallel corpus which is not surprising. In the opposite, we were expecting that for Twitter the results would be the worst, but it is not the case since it is extracted from Twitter by fixing the Hashtags. Consequently, both English and Arabic are related to the same topic which leads to the use of a narrow vocabulary. The worst results are achieved for the comparable corpora which is in this case the most difficult corpus for the coverage. In fact, it contains different documents related to several topics. This makes the vocabulary rich and disparate.
As presented before, the coverage rate has a serious impact in the calculation of Li and Gaussier comparability measure. That is why all the previous experiments have been presented. In Table 6, we present the comparability measure on the three previous corpora. The achieved results are correlated to the ones presented in the previous experiments. In fact, the best results concern the parallel corpus and the best method is the one which combines BAMA, LS and TreeTagger.

Table 6. The comparability measure of Li and Gaussier

Corpus	BAMA	LS	BAMA+ LS
Twitter	58.8	58.4	61
Euronews	49.6	49.9	51.6
Parallel	63.8	63.6	65

The comparability measure is not a final objective, we would like to extract for future work comparable tweets and comparable documents. That is why we conducted an experiment which looks for each tweet in Arabic its best corresponding tweet in English. In Table 7, we give respectively the maximum and the average value of comparability between a tweet in Arabic and the set of English tweets. The corresponding results are given also for a supposedly comparable corpus (Euronews) at respectively a sentence and document level. In this case, globally, the average comparability measure is better, at a document level for Euronews than for Twitter at a tweet level. In table 8, we give a sample of Arabic and their corresponding extracted English tweets *ExtTweet*

Table 7. Comparability values between Arabic and English documents

Corpus	Max	Average
Twitter	66	32.6
Euronews	60	27.3
Euronews document	65.8	35.8

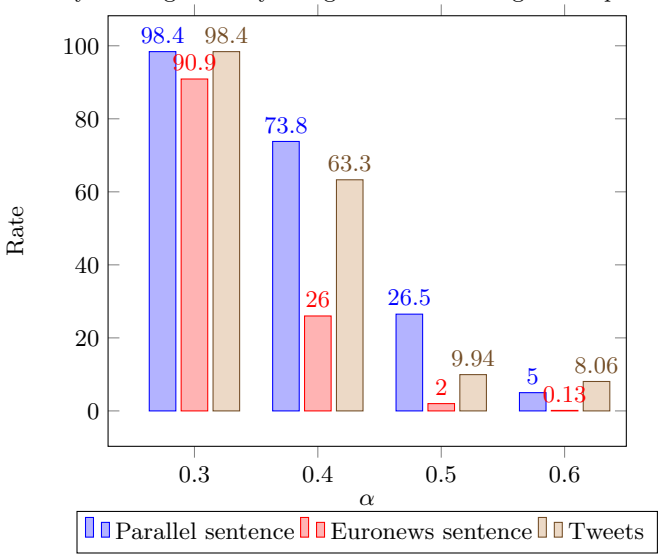
by using *LG* comparability measure. We give also the reference *Reference* which corresponds to the translation of the original tweet in Arabic.

In Figure 7, we give a chart describing the rate of pair of sentences for which the LG measure is greater than a certain threshold α for Twitter, Euronews and a Parallel corpus. This parameter could be considered such as a confidence measure that could be used in a future work to extract the best comparable and correlated tweets. We can notice that for $\alpha = 0.6$, we extract only 5% of comparable tweets, tweets supposed to handle the same idea. Only 8% from the parallel corpus are supposed to be comparable. This is an aberration, but this could be explained easily, in fact *LG* is based only on a bilingual dictionary while in parallel corpus a sentence is not obtained by a word by word translation of the other one. When we relax the constraints by accepting an *LG* greater than 0.3, we get for Twitter more than 98% of comparable tweets.

Table 8. A sample of Arabic and the extracted corresponding English tweets

Tweet	القيادة العامة للجيش تعلن السيطرة على المحطة الحرارية بريف حلب
ExtTweet	Syrian army has retaken completely under its control the thermal power plant east aleppo via
Reference	General Command of the Army announces the control of the thermal station in the countryside of Aleppo
Tweet	الهدنة الامريكية الروسية سوريا
ExtTweet	The american russian and iranian strategic triangle in syria
Reference	US Russian Truce Syria
Tweet	عاجل وقف اطلاق النار في سوريا لن ينجح ما لم تغير دمشق و موسكو تصرفاتهما الازمة السورية
ExtTweet	The american russian and iranian strategic triangle in syria
Reference	Urgent ceasefire in Syria will not succeed if Damascus and Moscow do not change their behavior in Syrian crisis

Fig. 7. The dictionary coverage rate by using all the stemming techniques on all corpora.



7 Conclusion and future work

In this paper, we investigate different multilingual corpora in order to make them comparable for a future work. For that, our experiments have been achieved on three Arabic-English corpora: Twitter, Euronews and ANN. The first one has been extracted by using specific Hashtags, the second one is supposed to be comparable and the third one is parallel. Processing data extracted from Twitter needs a cleaning and a harmonization of the data. That is why, we proposed several rules to make this corpus usable for the ongoing processes. To retrieve related documents from these corpora, we needed to measure their comparability. The similarity measure is the one proposed by Li and Gaussier which is based on a bilingual dictionary. Because, the coverage of the vocabulary is low, the Arabic text has been stemmed by BAMA and by a light stemming method, we proposed. Then we combined these methods in order to have a larger coverage. We applied also a stemming method to English part of Twitter by using TreeTagger. We expected that the coverage and consequently the comparability measure would be higher for Euronews than for Twitter, but it was not the case since the tweets are extracted by using specific Hashtags which makes the Arabic

and English tweets close. In addition, with this simple method we retrieved tweets which could be considered as comparable. This result will be used in a future work to retrieve parallel phrases to build a machine translation system based on social network data.

References

1. M. Attia, P. Pecina, A. Toral, L. Tounsi, and J. van Genabith. A lexical database for modern standard arabic interoperable with a finite state morphological transducer. In *Systems and Frameworks for Computational Morphology - Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings*, pages 98–118, 2011.
2. A. Balahur and M. Turchi. Improving sentiment analysis in twitter using multilingual machine translated data. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, pages 49–55, 2013.
3. T. Buckwalter. Buckwalter arabic morphological analyzer version 1.0. *Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.LDC2002L49*, 2002.
4. U. Bügel and A. Zielinski. Multilingual analysis of twitter news in support of mass emergency events. *IJISCRAM*, 5(1):77–85, 2013.
5. A. Cui, M. Zhang, Y. Liu, and S. Ma. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Information Retrieval Technology - 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings*, pages 238–249, 2011.
6. A. Fraisse and P. Paroubek. Twitter as a comparable corpus to build multilingual affective lexicons. *In The 7th Workshop on Building and Using Comparable Corpora, 26-31, May*, 2014.
7. P. Gamallo-Otero and I. Gonzalez-Lopez. Measuring comparability of multilingual corpora extracted from wikipedia. *Workshop ICL on Iberian Cross-Language NLP tasks., Huelva (Espanña), pages. 1-9.*, 2011.
8. G. Ke, P. F. Marteau, and G. Menier. Variations on quantitative comparability measures and their evaluations on synthetic french-english comparable corpora. *The International Conference on Language Resources and Evaluation*, 2014.
9. P. Koehn. Europarl: A parallel corpus for statistical machine translation. *Conference Proceedings: the tenth Machine Translation Summit, pages 79–86. Phuket, Thailand, AAMT*, 2005.
10. B. Li. *Measuring and improving comparable corpus quality.* PhD thesis, Université de Grenoble, 2014.
11. B. Li and É. Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 644–652, 2010.
12. E. Refaee and V. Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. *9th International Conference on Language Resources and Evaluation ,2014*, 2014.
13. S. Roukos, D. Graff, and D. Melamed. Hansard french/english. *LDC Catalog No:LDC95T20, 1995.*, 1995.
14. M. Saad, D. Langlois, and K. Smaili. Extracting comparable articles from wikipedia and measuring their comparabilities. *In V International Conference on Corpus Linguistics University of Alicante Spain*, 2013.
15. M. Saad, D. Langlois, and K. Smaili. Cross-lingual semantic similarity measure for comparable articles. *In Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings, pages 105-115. Springer International Publishing*, 2014.
16. L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, and L. Yi. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 1837–1842, 2014.