HAL
open science

# How to match bilingual tweets ?

Karima Abidi, Kamel Smaili

## To cite this version:

Karima Abidi, Kamel Smaili. How to match bilingual tweets ? . 6th NLP 2017 - Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT) , Feb 2017, Sydney, Australia. hal-01536078

**HAL Id: hal-01536078**

**https://hal.science/hal-01536078v1**

Submitted on 10 Jun 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.

# HOW TO MATCH BILINGUAL TWEETS?

Karima Abidi[1] Kamel Smaili[2]

[1] Ecole Supérieure d'Informatique (ESI), Algiers,Algeria
k_abidi@esi.dz
[2] Campus Scientifique LORIA , Nancy, France
kamel.smaili@loria.fr

***ABSTRACT***

*In this paper, we propose a method that aligns comparable bilingual tweets which, not only takes into account the specificity of a Tweet, but treats also proper names, dates and numbers in two different languages. This permits to retrieve more relevant target tweets. The process of matching proper names between Arabic and English is a difficult task, because these two languages use different scripts. For that, we used an approach which projects the sounds of an English proper name into Arabic and aligns it with the most appropriate proper name. We evaluated the method with a classical measure and compared it to the one we developed. The experiments have been achieved on two parallel corpora and shows that our measure outperforms the baseline by 5.6% at R@1 recall.*

***KEYWORDS***

*Comparability measure ; Arabic stemming, Proper names; Soundex, Twitter*

## 1. INTRODUCTION

The parallel corpora are extremely valuable resources for many applications in natural language processing, in particular for machine translation which needs massive corpus to train statistical models. However, this type of data are not always available and especially for certain pairs of languages. An attractive option is to collect data from the web and the social networks for which, nowadays data are abundant. However, this task is not easy to handle because the collected documents have to be aligned in accordance to their topic. When the corpora are aligned, they are considered as comparable. These corpora consist of a set of documents expressed in several languages which are not parallel in the strict sense, but deal with analogous subjects. These last decades this issue has grown considerably [10] [6] [13][16]. The community of Cross-Lingual Information Retrieval contributed significantly to propose different solutions to this issue [12], [7][14].

These last couple of years, a particular attention has been given to harvest data from social networks and particularly from Twitter. This is due to the fact that, so many people through the world adopted this social network to express their opinions. Consequently, it would be useful to investigate this media, in order to collect cross-lingual posts and align them in terms of topics. This will lead to get comparable corpora of Tweets. These documents could be then, used to extract parallel fragments or phrases. In order to have relevant parallel fragments, we need relevant comparable Tweets, to achieve that, we should propose an efficient measure to estimate the comparability. In this work, we will show that the classical dictionary-based measure [8] cannot be used directly for this task and especially for Arab Tweets. Specific knowledge related to Arabic will be taken into account in order to tackle, not only, the issue of the specificity of Arabic, but also this free way of writing the posts.

This paper is organized as follows: In section 2, we present related works concerning the extraction of comparable corpora. Then, we describe the corpus, we collect, and we give details about the preprocessing steps in sections 3 and 4. In section 5, we present the process we use

for matching bilingual Tweets. Several experiments and results are described in section 6. Finally, we conclude and present some future works.

## 2. RELATED WORKS

The construction of comparable corpora is performed using similarity measures. These measures can be based on three different approaches: vocabulary overlapping, vector space and Machine translation. Among existing work to align the comparable corpus, we can mention the following: Li and Gaussier in [8] 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 two parallel corpora to which noises have been added. They showed that the comparability degree decreased proportionally with the added noises. A similar approach proposed by Etchegoyhen et al. in [5] termed STACC, is based on expanded lexical sets and Jaccard similarity coefficient. The idea is to get rid of a manual bilingual dictionary. The bilingual dictionary is built on a large parallel corpus by using Giza ++[11]. Since, it is independent from languages, the approach has been evaluated on a large dataset of ten languages. Zhu et al. in [16] utilized a bilingual LDA model to match similar documents. They proved that this approach can obtain similar documents with consistent topics. Huang et al. in [6] describe a method based on techniques inspired from Cross Lingual Information Retrieval. With the translation of the keywords of the source documents, they retrieve the target documents which contain these translated words. Then, the mapping between source and target documents is achieved in accordance to a similarity value. A method based on word embedding has been proposed by Vulic et al in [14]. The model has the possibility to learn bilingual word embeddings from already comparable corpora. The crucial idea in this work is the fact that the method allows to share the cross-lingual embedding space.

Works on comparable corpus containing Arabic are not as popular as those used for English or French, we can mention those proposed in [10],[1], [12]. In this last work, different comparability measures based on bilingual dictionaries or on numerical methods such as Latent Semantic Indexing (LSI) have been proposed.

## 3. EXPERIMENTAL MATERIAL

As presented before, we propose to identify comparable corpora by extracting them from Twitter. In the following, we will be interested by two languages: Arabic and English. In order to identify comparable Tweets, we decided to set the topic which will be used to crawl the Tweets: *Syria's war*. Our objective is to crawl all the Tweets related to this subject and then to align Arabic and English at the Tweet level. For this purpose, we selected the 7th-top English Hashtags concerning the war in Syria (Table 1). The same process has been done on Arabic (Table 2).

Due to the particularity of the free way to write Arabic in Twitter, we used different Hashtags such as: #سوريا #سوريه , both of them correspond to the word *Syria*. One ending with *Ta* (ة) and the other with *alif* (ا).

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

| English Hashtag | $N_{Twts}$ |
|---|---|
| #SyrianRefugees | 10895 |
| #refugeescrisis | 2856 |
| #Syrianarmy | 3211 |
| #freesyrianarmy | 3119 |

| | |
|---|---|
| #SyriaCrisis | 6260 |
| #syria | 17000 |
| #syrian | 17000 |
| **Total** | **57771** |

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

| English Hashtag | | $N_{Twts}$ |
|---|---|---|
| #اطفال-سورية | Children of Syria | 1599 |
| #الثورة-السورية | Syrian revolution | 10092 |
| #اللاجئيين-السوريين | Syrian refugees | 4000 |
| #سوريا | Syria | 17000 |
| #سورية | Syria | 17000 |
| #الجيش-العربي-السوري | Syrian Arab army | 916 |
| #الجيش-السوري-الحر | Free Syrian army | 4318 |
| **Total** | | **59452** |

Table 1 and Table 2 show that the global number of crawled Tweets is approximatively equivalent. But the number of Tweets concerning the *Syrian Refugees* is not the same. It is twice much more in English than in Arabic. This could be explained by the fact that this topic has been very popular in the West, since the corresponding countries were directly concerned by the problem. While, the number of Tweets concerning *Syria* is much more important in Arabic than in English, since the Arab world is very involved in the Syrian issue.

## 4. PREPROCESSING TWEETS

In natural language and especially for processing Tweets, one needs to rewrite some words, to clean some of them, to homogeneous the way of writing, to transform digits, proper names, cities and so on. This step is referred as language preprocessing. In the following, we pre-process both Arabic and English by taking into account the linguistic specificity of each of them. This step is very crucial since our objective is to identify comparable Tweets. More the treatments of homogenization of Tweets in both languages are precise, and more the process of identifying comparable Tweets is relevant. Figure 1 illustrates an example of the differences between an English and its equivalent Tweet in Arabic. The fragments on red and green correspond to respectively the way to write the date and the number in Arabic.

| | |
|---|---|
| **English tweet** | Syria issues decree no 63 the general parliamentary elections will be on Wednesday april th 13 2016 |
| **Arabic tweet** | اصدر بشار الاسد المرسوم رقم ٦٣ القاضي بتحديد يوم الاربعاء ١٣ نيسان ٢٠١٦ موعد لانتخاب الاعضاء |

Figure 1. Example of comparable Tweets.

### 4.1. Preprocessing English Tweets

In the following, we present the main treatments we achieve on the English Tweet corpus. Since Twitter is used by hundred million of users and because a Tweet is limited to 140 characters, people take some freedom in writing their posts, for instance by shortening the words. To

handle this issue, we use a SMS dictionary [1] which contains abbreviations, acronyms and their literal corresponding text. We used this dictionary to replace abbreviations by their literal forms. For example, *ppl* will be replaced by *people*.

Sometimes, in an English Tweet, we can find some references to foreign languages. This could be a serious problem for the further linguistic treatments. Based on a list of stop words of few languages, we discard all the fragments which contain at least one of these stops words.

Contrary to Arabic Tweets corpus, the English one contains many Hashtags embedded in the Tweet itself. Sometimes they are in the middle of a post, consequently we cannot just remove them, since they are used as any word. In some Tweets, a Hashtag might be composed of several Hashtags *#SyrianArabArmy*, *#SyrianCivilians*, ...etc which should be split. In the case where words are separated by capital letters or special characters, it is easy to determine the border of words composing the compound Hashtag. But, when the Hashtag is completely written in lowercase, we need to perform differently. To do so, we decided to use a dictionary sorted alphabetically and by the word's length. We seek a word in the compound Hashtag by looking for it in this dictionary.

Because there are several ways to write a date in English, all of them will be homogenized such as: *DD/MM/Year*. In Table 3, we give some examples before and after the rewriting process.

In social network, a letter of some words could be duplicated several times to express an emotion or a sentiment, phenomenon known as an elongation such as in: *woahhh*. These kind of words are transformed by removing the duplicated letters: *woah*

Table 3. Examples of rewritten dates.

| Before | After |
|---|---|
| April 13,2016 | 13/4/2016 |
| february 1,2016 | 1/2/2016 |
| 22 feb 2016 | 22/2/2016 |
| 2.2.2016 | 2/2/2016 |

## 4.2. Preprocessing Arabic Tweets

Arabic is very different from Indo-European languages, that is why specific treatments have to be achieved, in order to make Tweets ready for the further processings.

For reasons of freedom writing in Twitter, users replace the letter ة by ه only at the end of words. This could be very surprising since these two letters have different linguistic roles in Arabic, but graphically they are similar, except that the first one has two dots above. For convenience, some people use indifferent one or the other. This is why we homogenized the script in all the tweets. For almost the same reasons, we replaced all the forms of symbol *Alif* with *Hamza* such as in: آ, أ, إ with a simple *Alif* ا . Concerning the diacritics, we removed them since people can read without short vowels.

In Arabic, two sets of digits are used for writing numbers, the first set is the one used around the world, known by Arabic digits and the second is the Indo-Arab digits which are much more used in Middle-East: ٠١٢..٩. For this purpose, we decided to keep only one numeral notation for

[1] www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm

numbers by using the English coma for decimal number, for example ٥,١٣will be rewritten as 5.13.

Concerning the dates, in the Arab world, three types of calendars could be used: Assyrian, Hijri and Gregorian. The first and the second one are much more used in the East than in the West of the Arab world. For example: *January* could be written: كانون الثاني , جانفي or يناير depending on the Arab region. That is why each date, whatever its form, is rewritten in accordance to the following pattern *DD/MM/Year* (see examples in Table 4).

In social networks, sometimes users stretch words to accentuate their opinion or just write the words such as they pronounce them, then it is necessary to normalize the way of writing these words. The stretched or duplicated letters are removed such as in the following Arabic examples:عاااجل → عاجل and ياااارب → يارب .

Table 4. Examples of date before and after the rewriting treatment.

| Before | After |
|---|---|
| الاول من شباط 2016 | 1/2/2016 |
| 22 فبراير 2016 | 22/2/2016 |
| 2.2.2016 | 2/2/2016 |
| 13,5 | 13.5 |

## 4.3. Stemming

Arabic language is morphologically rich due to the fundamental rules used for building the words. In fact, often a root is considered as a producer of words, since it is agglutinated to affixes and suffixes to form new words. For example: the root كتب ( *to write*), with specific affixes produces different words with different meanings: يكتب (*he writes*), مكتبة ( *library*), مكتب (*office*), etc.

To use statistical methods, the words have to be segmented in order to reduce the number of entries in the vocabulary and then to have relevant statistics. That is why, a stemming procedure is run in order to segment the Tweets. The idea is to replace different words which share the same root by the root itself. This will reduce the size of the list of distinct words and leads to a better coverage of the corpus.

In this work, we applied different techniques to retrieve the most representative form of Arabic words. To do so, we combined Buckwalter Arabic Morphological Analyzer with a method based on Light Stemming (LS) presented in previous work [2].

For English, we also used a morphological analyzer to reduce the flexional forms of English words, for that we used the TreeTagger tool [2].

## 4.4. Proper names in Arabic

Generally, detection of proper names is a critical task for natural language processing applications especially for Arabic. The problem becomes harder when the processing concerns the Tweets. For Latin languages, proper names start with a capital letter, unfortunately, for

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Arabic, this notion does not exist. In addition, personal names refer, in general to common words. For example, the first name كريمة means *generous*.

Another issue is that the proper names could be agglutinated to prefixes and suffixes which requires lemmatization step before performing identification of proper names. For example: بسوريا ( *in syria*) must be lemmatized بـ سوريا (Syria).

The majority of research work concerning the extraction of proper names have been dedicated to Modern Standard Arabic (MSA) by Pos-tagging, while only little attention has been given to unstructured text like tweets [15].

Our goal is how to match the English and Arabic proper names, from two bilingual Tweets. Name matching can be defined such as the process of determining, whether two name strings are instances of the same name [4]. This task is not difficult, if the two languages use the same alphabet. Otherwise, a transliteration of a source proper name has to be performed. Transliteration is the action of representing the signs of an alphabet of the source language by the signs of the target language. Transliterating Arabic is more difficult, because for each proper name, several acceptable transliteration candidates could be proposed depending on how it is pronounced in the target language. For example, for the first name: سليمان, the following transliterations are possible: *Sulayman, Seleiman, Sliman and Selayman*.

## 5. IDENTIFICATION OF COMPARABLE TWEETS

In general building comparable corpora consists in collecting multilingual documents concerning or not a specific topic, then documents are aligned by estimating the degree of their closeness. When we would like to build a comparable corpus of Tweets, the task is a bit complicated because the shortness of the post which makes the matching process more difficult. In fact, the matching process is based on the number of common words between two bilingual documents, unfortunately the number of words, in a Tweet, is very weak which makes this task harder. In the following, we propose a method dedicated to the extraction of comparable Tweets. From two Twitter corpus S and T, in two different languages and for a post $s_i^d$ published at date d, we look for the Tweet $t_j^{\acute{d}}$ published at date $\acute{d}$ respecting the constraint $d-1 \leq \acute{d} \leq d+1$.

We hope that, with this constraint, we retrieve Tweets which concern the same topic. In order to align Tweets, other processing steps are necessary, they are described in the following.

### 5.1. Number and dates identification

As presented in section 4.2 and 4.1, the processing of dates and numbers is a crucial step allowing to identify similar dates and numbers in Arabic and English Tweets. An homogenization of these items is done in accordance to Tables 3 and 4.

### 5.2. Proper names identification

To identify Arabic proper names several treatments have been applied. In Arabic, proper names could be simple such as علي ( *Ali*) or compounded such as: ابن عبد الرحمان ( *ibn Abdul Rahman*) or علاء الدين ( *Alaa Aldine*). These compounded proper names are generally composed with a single proper name preceded or followed by particles given in Table 5. We decided to merge them to facilitate their transliteration. For instance, ابن عبد الرحمان ( ibn Abdul Rahman) is rewritten into ابن_عبد_الرحمان *ibn_Abdul_Rahman*.

To facilitate the process of matching the proper names, for each Arabic particle, several transliterations are proposed (see Table 5).

Table 5. Particles used in the compound proper names.

| **Particles** | **Arabic (English Transliteration)** |
|---|---|
| **Prefix** | ( ابن, بن) (ibn, bin, ben), عبد ( 3bd, abd), ابو(abu, abo, abou), بنت (bint, bent), ام (oum) |
| **Suffix** | الدين (eldin, aldin, uldin,eldin) |

For the compounded proper names, it is easy to identify them thanks to the previous particles. While, for the single proper names the task is more difficult. That is why we encode them by taking advantage from their phonetic form. The encoding is done in both English and Arabic Tweet. In the English Tweet, all the words which are not in the bilingual dictionary, are encoded. The hypothesis is that we might consider them as proper names. Then all the words of the Arabic Tweet are encoded. If two strings from respectively Arabic and English Tweet have the same code, we can conclude that, one is the transliteration of the other. For that, we used Soundex [3] which proposes to replace each letter by the index of a group of characters. Each group is constituted by the letters corresponding to almost the same sound class (see Table 6). The characters of Group 0, are ignored unless they appear in the first position of the supposed proper name. Encoding consists in keeping the first character without any change and the following are encoded in accordance to Table 6. Any supposed proper name will be represented by a letter followed by three digits. For example, encoding the proper name جميلة, will give three codes (Figure 2) corresponding to the possible transliterations: *Djamila*, *Jamila* or *Gamila*.

Table 6. Encoding Table of Soundex.

| **English character** | **Index** | **Arabic character** |
|---|---|---|
| A E H I O U W Y | 0 | ي و ه ع ح ا |
| B P F | 1 | ب ف |
| C S K G J Q X Z | 2 | ك ق غ ص ش س ز ج ح |
| D T | 3 | ط ظ ض ذ د ث ت |
| L | 4 | ل |
| M N | 5 | م ن |
| R | 6 | ر |

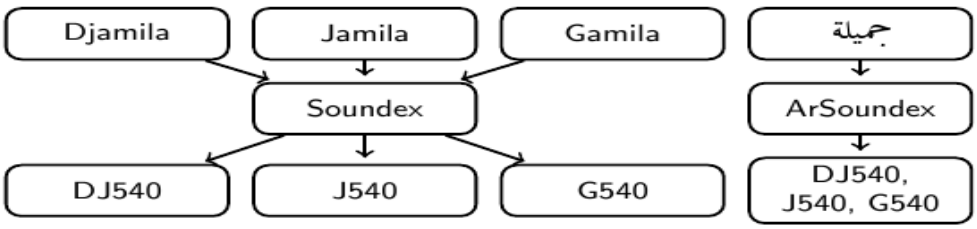


Figure 1. Encoding the proper name جميلة .

The weakness of Soundex is that, it encodes only the first four consonants. This constitutes a serious problem for the compound proper names. For example, *Abdel Aziz* will have the same code (A134) as *Abdel Rahman*. To overcome this constraint, we decided to encode each item of a compounded proper name. That is why, *Abdel Aziz* will be encoded by: *A134A220* and *Abdel Rahman* will be encoded by *A134R550*.

Comparing codes is not enough, in fact the encoding process does not produce a unique code for each proper name. So when an English word is encoded, it is then transformed into an Arabic word thanks to a transliteration table. Then the transliterated word is compared to those in the Arabic Tweet which have the same code as the English encoded word.

### 5.3. The comparability measure

In order to find similar Tweets, we used an adapted version of Li and Gaussier measure which is based on a bilingual dictionary [8]. The similitude between two Tweets is defined as follows: Let assume that T is an Arabic-English Tweet corpus consisting of an Arabic part $T_a$ and an English part $T_e$. The comparability measure can be defined as the maximum score between an Arabic Tweet $T_a$ and all the Tweets $T_i$ for $1 \leq i \leq N_{Te}$ Where $N_{Te}$ is the size of $T_e$.

For each Arabic Tweet, Score is calculated as follows:

$$Score(t_a, t_e) = \frac{\sum_{w \in T_e \cap D_e} \sigma(w, T_a) + \sum_{w \in T_a \cap D_a} \sigma(w, T_e)}{|T_e \cap D_e| + |T_a \cap D_a|} \quad (1)$$

where $D_e$ (respectively $D_a$) is the English side (respectively Arabic side) of a bilingual dictionary. $\sigma$ is a function which indicates if one of the potential translations T(w) of the word w does exist in the vocabulary $V_x$.

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

The adapted comparability measure of Li and Gaussier referred as LGT is calculated as follows:

$$LGT(t_a) = \max_{1 \leq i \leq N_{Te}} Score(t_a, t_e^i) \quad (3)$$

The size and quality of dictionary may heavily affect the result of comparability measure. In this work, we used the Open Multilingual WordNet (OMW) which contains 17 785 pairs of Arabic and English word. In the previous sections, we proposed several procedures in order to take into account the specificities of Arabic and particularities of the Tweets. In other words, when we identify a number, a date or a proper name, this has to be taken into account in the comparability measure. That is why, we modified the score mentioned in (1) as follows:

$$MScore(t_a, t_e) = \frac{\sum_{w \in T_e} \sigma(w, T_a) + \sum_{w \in T_a} \sigma(w, T_e)}{|T_e| + |T_a|} \quad (4)$$

$\sigma$ returns 1 if w has a translation in the target Tweet or if it has been identified such as a number, a date, a proper name, etc. (see section 5). And the new comparability measure is calculates as follows:

$$MLGT(t_a) = \max_{1 \leq i \leq N_{Te}} MScore(t_a, t_e^i) \quad (5)$$

## 6. EXPERIMENTAL RESULTS

The idea of identifying comparable corpora is an intermediate milestone, the final goal is to look for the best matched Tweets, in order to retrieve parallel phrases which could be used in machine translation. First of all, we need to measure the comparability of the collected bilingual corpus of Tweets by using the measures presented in Section 5.3. To evaluate the reliability of this measure we run an experiment on two parallel corpora: The first extracted from Twitter [9] which is available at [3] referred in the following as $C_{\text{Ling}}$. This corpus contains just 2006 parallel tweets. To our knowledge, it is the only available parallel corpus Arabic-English. Since this corpus is small, we decided to test on a parallel newspaper corpus which contains 11942 sentences extracted from ANN[4], referred in the following as $C_{\text{ANN}}$ .

Concerning the Twitter corpus of Ling et al., unfortunately when we investigated its content, we discovered that the data are not really parallel. Consequently, we cannot used it as a reference corpus for our tests. To illustrate the problems we encountered, we give in Table 7, few examples of what has been considered by the authors as parallel Tweets.

Table 7. Example of parallel tweets extracted by [9] .

| Nb | source and target tweet |
|---|---|
| 1 | **Source**: اعرف منين امال ماهر<br>**Target**: a3raf menen amal maher |
| 2 | **Source**: ★*○○○*★*★*○○○★★-★<br>**Target**: ★*○○○*★*★*○○○★★ |
| 3 | **Source**: $$---$$ الدراما هذي جمممال فيها المجدول الي ب 39 يوم<br>**Target**: $$,$$ |
| 4 | **Source**: قطع الماس الماس!!!<br>**Target**: diamond cut diamond !!.. diamante taglio |
| 5 | **Source**: تم العثور علي شخص فاهم خطاب مرسي جاري التحقيق معه<br>**Target**: c'est la vie |
| 6 | **Source**: ستار اكاديمي 9 في المسبح<br>**Target**: فضائح ستار اكاديمي9 |

In this corpus some tweets are considered as parallel, though they contain only characters such as in the 2 *th* and 3 *th* examples. This is due to the absence of preprocessing before alignment.

To identify the language of tweets, the authors used a binary function which yields 1 if a word w contains characters of a specific language L, and 0 otherwise. This function is useful if we would like to differentiate Mandarin from English, but not English from French. From a subset of 1000 Tweets, 0.6% of tweet are different from the desired language such as in the 4 *th* and 5 *th* examples.

We found in this parallel corpus, a Tweet in a language which is aligned with itself in the same language which is unacceptable for our purpose (see the 6 *th*).

[3] http://www.cs.cmu.edu/~lingwang/microtopia/
[4] www.annahar.com

Due to all these problems, we decided to select 1000 multilingual tweets considered by the authors as parallel and we extracted the real parallel Tweets. Only 34% are strictly parallel. The following tests have been achieved on this subset of parallel Tweets called$C_{Ling34}$.

We run several experiments with the two comparability measures described in Section 5.3 We calculated the classical Recall (R@1, R@5 and R@10). Results are presented in Table 8.

Table 8. LGT and MLGT results for parallel corpora.

| Corpora | Method | R@1 | R@5 | R@10 |
|---|---|---|---|---|
| $C_{ANN}$ | LGT | 73 | 85 | 87 |
| | MLGT | 79.7 | 89.4 | 92 |
| $C_{Ling34}$ | LGT | 53 | 73 | 78 |
| | MLGT | 56 | 77 | 83 |

This table shows that the modified LGT achieves better results than LGT since it takes into account the treatment of proper names, dates and numbers. For Twitter corpus, the recall is 56% at R@1 and grows up to 83% at R@10. This result is interesting, it allows to retrieve in the top-10 the right target Tweet.

Concerning the newspaper corpus which is larger, the results are more crucial since at Top-10, we get a recall of 92%. In Table 9, we give some examples of matched tweets.

Table 9. Example of comparable tweets aligned by LGT .

| |
|---|
| **Source**:minister kerry: the diplomatic path is the only path that can isolate terrorist groups like daash and front victory 1 2 syria .<br>**Target**: الوزير كيري المسار الدبلوماسي هو المسار الوحيد الذي يمكن ان يعزل الجامعات الارهابية مثل داعش وجبهة النصرة 1 2<br>**Translation**: minister kerry: the diplomatic path is the only path that can isolate terrorist groups like daash and front |
| **Source**: news: obama called putin on syria ceasefire: white house<br>**Target**: البيت الابيض اوباما وبوتن يبحثان وقف اطلاق النار في سوريا<br>**Translation**: white house: obama discusses with putin on syria ceasefire |
| **Source**: Syria president bashar al assad issues decree no 63 which sets wednesday 13/4/2016<br>**Target**: سورية اصدر الرئيس بشار الاسد المرسوم رقم 63 لعام 2016 القاضي بتحديد يوم الاربعاء 2016/4/13 موعدا<br>**Translation**: Syria president bashar al assad issues decree no 63 which sets wednesday 13/4/2016 |

## 7. CONCLUSION

In order to obtain a parallel Twitter corpus for which further NLP process could be considered, we developed a method which allows to align the Tweets of a same topic. The experiments achieved showed that for a Tweet, the proposed method can retrieve the corresponding target Tweet with a recall of 83% at R@10. This result has been achieved by a series of preprocessing which permits to align more easily two bilingual Tweets (Arabic and English). Preprocessing is a crucial step, when the data are extracted from social networks and more particularly from those which are written in Arabic. Specific treatments of dates, numbers and transliteration of proper names have been proposed to overcome the issues relates to this specificity. This

preprocessing allowed to improve the results by 5.6% in comparison to the baseline model. This result is very encouraging and will permit in a future work to consider the extraction of parallel phrases.

## REFERENCES


[1] Sadaf Abdul-Rauf & Holger Schwenk (2009) "On the Use of Comparable Corpora to Improve {SMT} performance", {EACL} 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009.

[2] Karima Abidi & Kamel Smaili, (2016), "Measuring the comparability of multilingual corpora extracted from Twitter and others", The Tenth International Conference on Natural Language Processing, HrTAL2016, Croatia, 29 September – 1 October 2016.

[3] Syed Uzair Aqeel & Steven M. Beitzel & Eric C. Jensen & David A. Grossman & Ophir Frieder (2006) "On the development of name search techniques for Arabic", JASIST .

[4] Peter Christen (2006), "A Comparison of Personal Name Matching: Techniques and Practical Issues", Workshops Proceedings of the 6th {IEEE} International Conference on Data Mining {(ICDM} 2006), 18-22 December 2006, Hong Kong, China.

[5] Thierry Etchegoyhen & Andoni Azpeitia, (2016) "Set-Theoretic Alignment for Comparable Corpora", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, {ACL} 2016, August 7-12, 2016, Berlin, Germany, Volume1: Long Papers.

[6] Degen Huang & Zhao, Lian & Li, Lishuang & Yu, Haitao(2010)," Mining Large-scale Comparable Corpora from Chinese-English News Collections", Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, Beijing, China

[7] Andrey Kutuzov & Mikhail Kopotev & Tatyana Sviridenko & Lyubov Ivanova, (2016), "Clustering Comparable Corpora of Russian and Ukrainian Academic Texts:Word Embeddings and Semantic Fingerprints", CoRR.

[8] Bo Li & Eric Gaussier (2010) , "Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora", COLING} 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China.

[9] Wang Ling & Guang Xiang & Chris Dyer & Alan W. Black & Isabel Trancoso (2013), "Microblogs as Parallel Corpora", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, {ACL} 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers

[10] Dragos Stefan Munteanu & Daniel Marcu (2005)," Improving Machine Translation Performance by Exploiting Non-Parallel Corpora", Computational Linguistics .

[11] Franz Josef Och and Hermann Ney (2003), "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, Page: 19--51.

[12] Motaz Saad and David Langlois and Kamel Smaili (2014), "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.

[13] Inguna Skadina & Ahmet Aker &Voula Giouli & Dan Tufis & Robert J. Gaizauskas & Madara Mierina &Nikos Mastropavlos (2010) ), "A Collection of Comparable Corpora for Under-resourced Languages", Human Language Technologies - The Baltic Perspective - Proceedings of the Fourth International Conference Baltic {HLT} 2010, Riga, Latvia, October 7-8, 2010.

[14] Ivan Vulic & Marie Francine Moens (2015), "Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings", Proceedings of the 38th International {ACM} {SIGIR} Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015.

[15] Omnia H. Zayed and Samhaa R and El-Beltagy,(2015) “Hybrid Approach for Extracting Arabic Persons Names and Resolving their Ambiguity from Twitter”, In 20th International Conference on Application of Natural Language to Information Systems (NLDB 2015), Passau, Germany, June. Springer.

[16] Zhu, Zede and Li, Miao and Chen, Lei and Yang, Zhenxin,(2013), “Building Comparable Corpora Based on Bilingual LDA Model.”, ACL (2) Page: 278--282.

**Author**

Photo