CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube
Karima Abidi, Mohamed Amine Menacer, Kamel Smaili

To cite this version:
Karima Abidi, Mohamed Amine Menacer, Kamel Smaili. CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube. 18th Annual Conference of the International Communication Association (Interspeech), Aug 2017, Stockholm, Sweden. hal-01531591

HAL Id: hal-01531591
https://hal.science/hal-01531591
Submitted on 1 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.
CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube

K. Abidi¹,², M.A. Menacer² and Kamel Smaili ²

¹École Supérieure d’Informatique (ESI), Algiers, Algeria
²SMarT Group, LORIA, F-54600, France

k_abidi@esi.dz, mohamed-amine.menacer@loria.fr, kamel.smaili@loria.fr

Abstract
This paper addresses the issue of comparability of comments extracted from Youtube. The comments concern spoken Algerian that could be either local Arabic, Modern Standard Arabic or French. This diversity of expression gives rise to a huge number of problems concerning the data processing. In this article, several methods of alignment will be proposed and tested. The method which permits to best align is Word2Vec-based approach that will be used iteratively. This recurrent call of Word2Vec allows us improve significantly the results of comparability. In fact, a dictionary-based approach leads to a Recall of 4, while our approach allows one to get a Recall of 33 at rank 1. Thanks to this approach, we built from Youtube CALYOU, a Comparable Corpus of the spoken Algerian.

Index Terms: Algerian dialect, Word2Vec, Comparable corpora, Soundex

1. Introduction
Nowadays, to develop speech recognition systems for vernacular languages, is a real challenge for the community. Contrary to classical natural language, vernacular languages are by definition used daily in the communication but are rarely written. To develop a speech recognition system or machine translation for this kind of language, we need to collect data. The best way to get this type of data is to harvest them from social networks. In this work, we are interested by the spoken Algerian. One needs to know that in Algeria people speak their mother tongue, which is a special Arabic dialect, but could speak the official language, which is Modern Standard Arabic (MSA), and also French. Nevertheless, what makes the issue more challenging is that people can mix in the same sentence the three previous languages.

Extracting data from Youtube is not new, Salama et al. in [1] extracted several dialects from this social network in order to constitute a dataset of several dialects. In [2], the authors used Youtube in order to build a Lexicon for the sentiment Analysis. A similar work has been done by [3]. In [4], the authors studied the influence of the sentiment existing in comments of Youtube. Our objective in this work is to collect from Youtube Algerian comments and to align those that are close to each other. To do that, we crawled 22000 videos by using 250 queries related to Algerian topics, such as: Bouteflika, Sellal, Messaoudi, Samira Tv, Allouet, etc. Aligning comments will produce a comparable corpus we called CALYOU (Comparable spoken Algerian extracted from YOUtube). This corpus could be used in several topics such as: adapting a MSA language model to Algerian dialect, to extract parallel dialectal phrases, which may be included in translation tables, etc.

This paper is organized as follows: In Section 2, we present related works concerning the extraction of comparable corpora. Then, we give an overview on languages spoken in Algeria and their difficulties, in Section 3. We describe the collected corpus in Section 4. Several experiments and results are described in Section 5. Finally, we conclude and present some future works.

2. Related works
The comparability topic is widely used in the community and several articles have been published. When documents are collected, one needs to measure the closeness between different articles. The automatic acquisition of comparable corpora can be accomplished by using similarity measures. These measures consist to quantify and capture the different comparability levels. Several related work proposed efficient measures to align and evaluate the quality of extracted comparable documents. Li and Gaussier in [5] 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 that estimates the comparability of two parallel corpora to which noise have been added. They showed that the comparability degree decreased proportionally with the added noise. A similar approach proposed by Etchegoyhen et al. in [6] 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 ++[7]. Since, it is independent from languages, the approach has been evaluated on a large dataset of ten languages. Huang et al. in [8] 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 that 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 Vulić et al in [9]. 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],[11], [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.

At the best of our knowledge, there is no work consisting of aligning corpora of the Algerian spoken language in its all forms: writing in MSA, in French and in dialect. Add to that, all the issues related to the way of writing the dialect: Latin script or Arabic script. Also, due to the missing of dictionaries and corpora, no standardization of writing words is available. Concerning resources, except PADIC [13] a multilingual parallel corpus, which contains 6 dialects including two from Algeria,
there is no substantial known data.

3. The spoken Algerian

People in Algeria speak mainly Algerian dialect, but this one could be mixed to Modern Standard Arabic and even French. Algerian dialect is informal language, not used in official speech, it includes an important portion of borrowed words, especially from French, Turkish and Berber. Vocabulary of Algerian dialect includes verbs, nouns, pronouns and particles. Most of this vocabulary is from MSA. Moreover, there is significant variation in the vocalization in most cases, and omission of some letters in other cases (mainly the Hamza)\(^\dagger\). As presented before, in this paper we are interested by the comments of Algerian videos. We have to mention that there is no rule in the way of writing the comments and this constitutes a real challenge for the community. Because no standardization exists, people write sometimes Algerian dialect by using Latin script (LS), called also Arabizi that is an alphabet to communicate over Internet. This obviously constitutes a serious issue because Arabic NLP tools could not be used. Unfortunately, French NLP tools cannot be used either since the written words in LS are not in French. In the following, we give an example of a post related to an Algerian video:

 '\'w jma3atkom t3ayeb f bouteflika aw rah khirmanek b alf khatra.\nThat should be written with Arabic script (AS) such as:\n\(" و هما شتمك تعيّب في بوططفة أو راه خير مند بلاف خطرة\) and means: And you are making fun of Bouteflika when he is a thousand times better than you.\n
One can remark that people use some codification for specific Arabic letters that do not exist in LS. This is the case of \(\text{؟} \) that is written \(\text{p} \), that is written \(\text{9} \), \(\text{j} \) as \(\text{s} \) and other codes that are not officially unfortunately adopted by everyone.

Algerian dialect could be written by using AS, but some words are not Arabic, they are in general adapted from French such as:\n\(" بالأسماء \) that means its place. The sound / p / does not exist in Arabic, but in Algerian dialect it is used as it is and sometimes it is replaced by / b /.

In addition to the Algerian mother tongue, people may express their opinions in French, while some propose their comments in Modern Standard Arabic. This diversity of the different ways to write comments show the large panel of writing the spoken language that make the processing very challenging.

4. Corpus

To build CALYOU, we harvested data from Youtube by using the Google’s API\(^2\) that allows users to search for videos that match specific criteria and retrieve all information and comments of those videos. To harvest, we chose few keywords to form queries in order to retrieve videos concerning national news, Algerian celebrity, football, etc. Table 1 shows some statistics of the collected data, where |C| is the number of comments before and after preprocessing, |W| is the number of words and |V| is the vocabulary size.

We can mention that after the cleaning process, the corpus has been reduced by around 20% and the vocabulary by around 10%.

5. Measuring the comparability

Our objective is to make comparable Algerian dialect, French and Arabic comments to create CALYOU. What we propose, is not only, to produce comparable French and dialect comments, but also to align dialect comments written in LS with those written in AS.

To this end, we need to measure the comparability of comments, that is why we used classical approaches and proposed new ones adapted to the Algerian dialect.

The methods are evaluated on a test corpus composed of comparable comments built by hand. The performance are given in terms of the classical Recall (R@1, R@5 and R@10), well known in Information Retrieval, are given in Table 3 and will be detailed for each approach in the next sections.

5.1. Dictionary-based method

In a first experiment, we used a dictionary-based method proposed by Li and Gaussier (LG)\(^5\). The comparability measure can be defined as the maximum score between a comment written in LS and all the comments written AS. We mention that even if the comment is written in LS, that does not mean necessarily that it is a French comment. It could be a dialectal comment written in LS as explained in Section 3.

The size and the quality of dictionary may heavily affect the result of comparability measure. For LG, we used the Open Multilingual WordNet (OMW)\(^3\) that contains 55373 pairs of Arabic and French words. LG method achieves bad results: 4 at Recall 1 (see Table 3), this was expectable since the dictionary is composed by MSA words and not by Algerian Arabic words. In our best of knowledge, there is no available large Algerian - French dictionary, that is why we used the mentioned dictionary because even if the comments are in Arabic dialect, they contain words in MSA.

5.2. Indexing words by their sounds

Since dialect is a spoken language, people may write the same word in different manners, while keeping the same pronunciation. That is why, we decided to improve the last method by indexing the words by their pronunciation or more exactly by their representative sounds. For that, we used Soundex\(^1\), a phonetic algorithm for indexing by sound. Words are encoded by taking advantage of their phonetic form. The encoding is done in both comments in LS and AS. All the words in the LS that do not exist in the bilingual dictionary are encoded by Soundex. The underlying hypothesis is that we might consider them such as dialectal words or proper names. Then all the words of the comment written in Arabic are encoded by Soundex. If two strings from respectively LS, AS comments have the same code, we can conclude that, one is the transliteration of the other. Soundex proposes to

\(^{1}\)The Hamza is a letter in the Arabic alphabet, representing the glottal stop.
\(^{2}\)Available at: https://developers.google.com/Youtube
\(^{3}\)Available at: http://compling.hss.ntu.edu.sg/omw/
replace each letter by the index of a group of characters. Each group is constituted by the graphemes corresponding to the similar class of sounds (see Table 2). We obviously adapted the original corresponding’s table in order to match the graphemes of Youtube comments and Arabic sounds. The characters of Group 0, are ignored unless they appear in the first position of a word. Encoding consists in keeping the first character without any change and the following are encoded in accordance to Table 2. Any word will be represented by a letter followed by three digits. For example, encoding the dialect word 7ouma will give two codes (Figure 1) corresponding to the possible transliterations: Houma and 7ouma.

Table 2: Corresponding’s table for transliteration

<table>
<thead>
<tr>
<th>English character</th>
<th>Index</th>
<th>Arabic character</th>
</tr>
</thead>
<tbody>
<tr>
<td>A E H I O U W Y</td>
<td>0</td>
<td>ى و د ع ا</td>
</tr>
<tr>
<td>B P F</td>
<td>1</td>
<td>ب ف</td>
</tr>
<tr>
<td>C S K G J Q X Z</td>
<td>2</td>
<td>ل ت</td>
</tr>
<tr>
<td>D T</td>
<td>3</td>
<td>و ن</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
<td>ر</td>
</tr>
<tr>
<td>M N</td>
<td>5</td>
<td>ن م</td>
</tr>
<tr>
<td>R</td>
<td>6</td>
<td>ه</td>
</tr>
</tbody>
</table>

Table 3: Comparability results

<table>
<thead>
<tr>
<th></th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>4</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>LG + Sound</td>
<td>11</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td>SCw2v</td>
<td>23</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td>SCw2v + Sound</td>
<td>24</td>
<td>45</td>
<td>52</td>
</tr>
<tr>
<td>Iterative SCw2v + Sound</td>
<td>33</td>
<td>48</td>
<td>54</td>
</tr>
</tbody>
</table>

5.3. A Word2Vec-based method

The free style of writing in the comments of Youtube gives rise to some problems when words should be matched. In fact, the same word in the Algerian spoken could be written in several ways such as in Table 4. The word kho, which means brother, could be written in various manners in addition to similar words such as hbibna. The word fr is also a way to express the meaning of the word kho; it is derived from the French word frère of which the vowels have been discarded. Because we work on dialects and there is no dictionary, which could provide all the forms related to one lexical entry, we decided to use Word2Vec to retrieve the different forms of a word. To find the correlated words for each entry of an Algerian lexical dialect, we opted for a continuous bag of words (CBOW) method [15] with a sliding window of 150. This size has been fixed after several tests. This large number is due to the fact that all the comments concerning the same video have been concatenated into one document. CBOW allows us to build for each word, a list of its correlated words, the idea is to have for each word all its various forms. The comparability measure σ(w,C) = w,C where σ is the function, which increases the value of the comparability, if any word provided by CBOW does exist in the target comment.

Table 4: An Example of different ways to write the word kho

<table>
<thead>
<tr>
<th>khoya</th>
<th>kho</th>
<th>kho</th>
<th>kho</th>
<th>kho</th>
<th>kho</th>
<th>frr</th>
</tr>
</thead>
<tbody>
<tr>
<td>khoo</td>
<td>hbibna</td>
<td>kho</td>
<td>kho</td>
<td>fr</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: The closest words of kho in accordance to CBOW method

<table>
<thead>
<tr>
<th>khoya</th>
<th>video</th>
<th>kho</th>
<th>khoya</th>
<th>rak</th>
<th>kho</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.59</td>
<td>0.53</td>
<td>0.52</td>
<td>0.51</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>khoo</td>
<td>ta3</td>
<td>hbibna</td>
<td>kho</td>
<td>kho</td>
<td>fr</td>
</tr>
<tr>
<td>0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.41</td>
<td>0.38</td>
<td>0.37</td>
</tr>
</tbody>
</table>

by Word2Vec to better estimate the comparability between two comments. The comparability between a comment in LS with a comment in AS is estimated as follows:

\[ SC(C_{LS}, C_{AS}) = \sum_{w \in C_{LS}} \sigma(w, C_{AS}) + \sum_{w \in C_{AS}} \sigma(w, C_{LS}) |C_{LS}| + |C_{AS}| \]

where \( C_{LS} \) (respectively \( C_{AS} \)) is Latin Script (respectively Arabic script) of Algerian video comments. \( \sigma(w, C) \) is a function, which increases the value of the comparability, if any word provided by CBOW does exist in the target comment.

\[ \sigma(w, C) = \begin{cases} 1 & \text{if } \exists m/m \in C \wedge m \in C \\ 0 & \text{else} \end{cases} \]

The comparability measure \( SC_{w2v} \) can be defined as the maximum score between a Latin script comment \( C_{LS} \) with every comment written in Arabic script \( C_{AS} \) for \( 1 \leq i \leq N_{C_{AS}} \). Where \( N_{C_{AS}} \) is the number of comments written in Arabic script.
In order to improve the comparability of comments, we decided to iterate the process of 
Word2Vec. The underlying idea is to improve, at each step, the quality of comparability of the 
comments. In fact, since the results of comparability have been improved by using Word2Vec, we decided to retrain the system with better comparable documents achieved by this method. That means, after getting comparable comments by using Word2Vec, these documents are put back into the training corpus. Then, we calculate again for each word in Latin script its corresponding and correlated words written in Arabic script. The justification of this process is that at the initial step, the CBOW method runs on articles composed of bulk comments, but of the same topic. Theses comments are not gathered in a training corpus. Then this has been exploited for the iterative process of comparability. This iterative process is repeated until the measure of comparability stops increasing as illustrated in Figure 2.

This process has been iterated several times until the Recall 1 stops increasing. In fact, this process improves the quality of comparability that reaches 33 at Recall 1 and 54 at Recall 10. This improvement is significant in comparison to all the other experiments achieved before.

In Table 6, we present some examples of comments aligned by our approach. For each comment in Latin script, the corresponding target comment retrieved by the iterative Word2Vec is given. One can notice that, in the first example of Table 6, the French comment is not well written. As presented before, people write sometimes as they want without any respect of the linguistic rules. We can notice in these examples that the retrieved comments are close to the source comments. Let us focus on the fact that sometimes the dialect is written in Latin script but the sentence is not in French, such as in the third example of Table 6. Our approach permits to find the best comment written in Arabic associated to this dialect comment written in Latin script.

Table 6: Example of comparable comments aligned by using the iterative Word2Vec principle

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Target</td>
<td>Translation</td>
</tr>
<tr>
<td>Source</td>
<td>Target</td>
<td>Translation</td>
</tr>
</tbody>
</table>

6. Conclusions

In this article, we addressed the difficult issue of matching comments from Youtube for a vernacular language for which no writing rules do exist. It concerns the spoken Algerian which people express their opinions in Youtube in local Arabic dialect, Modern Standard Arabic and French. Sometimes, in the same comment the three languages may be mixed. This particularity is specific to the spoken Algerian that leads to more difficulties in the treatments of the corresponding texts. We tested a classical dictionary-based method to match comments; this approach totally collapsed by achieving a Recall of 41. Because, people write with Latin script and Arabic script, the use of phonetic representation has been considered in order to find the transliterations of words written in two completely different scripts. The use of a phonetic algorithm (Soundex) allowed to reach an absolute improvement of 7 points. This improvement is relevant, in comparison to the initial score of Li and Gaussier that is very poor. Then, we used a Word2Vec-based approach to find a list of words that could be correlated to a lexical entry. This method permitted to find a list of variations of the same word. Then this has been exploited for the matching process of comparability. The results jumped again and outperformed the baseline method. We proposed then an approach that iterates the process of collecting the list of words by Word2Vec in order to improve at each step the quality of the training corpus. The results achieved are very relevant since the corresponding method outperforms the approach LG and Soundex by 22 points. This approach allowed us to build CALYOU: A comparable spoken Algerian corpus extracted from YouTube. The performance we get (33 at Recall 1) may be considered as weak and we have to improve it in a future work. But, different experiments in the literature, on comparable corpora for natural language and not for dialects, have been done, but they have not reached much better results [16] than what we get on much more difficult data.
7. References


