INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube

K. Abidi[1,2]*, M.A. Menacer*[2] *and Kamel Smaïli* [2]

[1]École Supérieure d'Informatique (ESI), Algiers, Algeria
[2]SMarT 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, Aliouet, 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 Vulič 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,

http://dx.doi.org/10.21437/Interspeech.2017-1305

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)[1]. 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.
That should be written with Arabic script (AS) such as:
و جماعتكم تعيب في بوتفليقة او راه خير منك بالف خطرة
and means: *And you are making fun of Bouteflika when he is a thousand times better than you.*
One can remark that people use some codification for specific Arabic letters that do not exist in LS. This is the case of ع that is written *3*, ق that is written *9*, خ as *5* 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: بلاصتو 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%.

[1]The Hamza is a letter in the Arabic alphabet, representing the glottal stop.

[2]Available at: https://developers.google.com/Youtube

Table 1: *The collected Youtube Algerian Dialect Corpus*

	Raw corpus	Cleaned Corpus
$\|C\|$	1.1M	853K
$\|W\|$	16M	12.7M
$\|V\|$	97K	88K

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 [14], 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

[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 حومة, will give two codes (Figure 1) corresponding to the possible transliterations: *Houma* and *7ouma*.

Table 2: *Corresponding's table for transliteration*

English character	Index	Arabic character
A E H I O U W Y	0	ى 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	ظ ط ض ذ د ث ت
L	4	ل
M N	5	م ن
R	6	ر

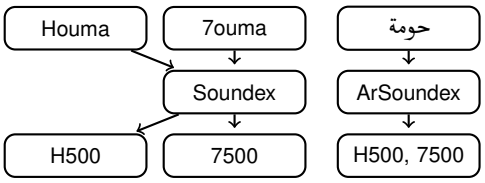

Figure 1: *Encoding dialect word* حومة

For the previous example حومة will be encoded by "ح500", then ح is transliterated by *7* or *H*. Consequently, the words حومة, *Houma* and *7ouma* match. Therefore, the process of comparability will provide better results. In fact, we combined LG and Soundex in the process of comparability. The results are given in Table 3 (Line 2). Introducing Soundex allows to improve the results from 4 to 11 at Recall 1. This is obviously not enough but, henceforth the use of the sound representation of words is necessary.

Table 3: *Comparability results*

	R@1	R@5	R@10
LG	4	12	19
$LG + Sound$	11	27	40
SC_{w2v}	23	42	51
$SC_{w2v} + Sound$	24	45	52
$Iterative\ SC_{w2v} + Sound$	33	48	54

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 *frr* 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

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

khoya	khoo	khouya	khou	khyo
khooo	hbibna	خويا	خو	frr

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. Table 5 shows the words that are related to *kho* with the corresponding distances. We can remark that the majority of those words of Table 4 do exist in the list provided by word2Vec. In addition to these words, others are proposed that will be helpful for the comparability. We used these lists of words provided

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

khoya	video	khoo	khouya	rak	khou
0.59	0.53	0.52	0.51	0.47	0.46
khooo	ta3	hbibna	خويا	خو	frr
0.45	0.44	0.44	0.41	0.38	0.37

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:

$$SC(C_{LS}, C_{AS}) = \frac{\sum_{w \in C_{LS}} \sigma(w, C_{AS}) + \sum_{w \in C_{AS}} \sigma(w, C_{LS})}{|C_{LS}| + |C_{AS}|} \quad (1)$$

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

$$\sigma(w, C) = \begin{cases} 1\ if\ \exists m/m \in CBOW(w) \wedge m \in C \\ 0\ else \end{cases} \quad (2)$$

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^i_{AS} for $1 \leq i \leq N_{C_{AS}}$. Where $N_{C_{AS}}$ is the number of comments written in Arabic script.

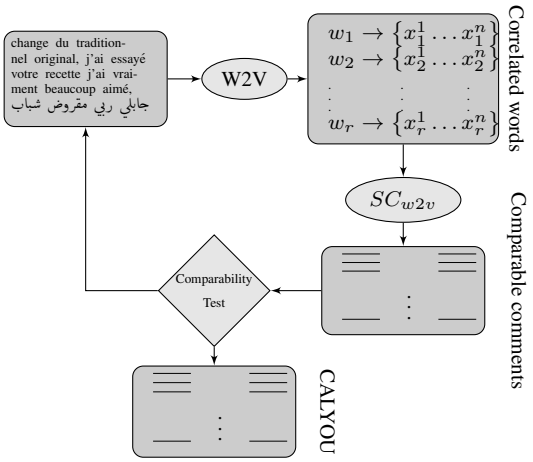

Figure 2: *Iterative process of comparability based on Word2Vec*

$$SC_{w2v}(C_{LS}) = \max_{1 \leq i \leq N_{C_{AS}}} Score(C_{LS}, C_{AS}^i) \qquad (3)$$

The results of comparability are better than what we obtain with Li and Gaussier combined with Soundex. In fact, in this method we take into account the variability of writing the dialect in the comments of Youtube videos. The results jumped from 11 at Recall 1 to 24, this is a significant result but is not enough. By combining the comparability based on CBOW with Soundex, we noticed a small improvement of 4% at rank 1 and 6% at rank 5 (see Table 3 Line 4) .

5.4. Iterating Word2Vec for improving the comparability

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 order to assure good performance to CBOW. When we align them, we hope that by injecting them into the training process, we will get a better training corpus. 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 theses 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*

Source j'ai trop aimé la tenu c mon style **Translation** I like too much your outfit, this is my style **Target** عجبوني بزاف نحب هاد ستيل **Translation** I like them too much, I like this style
Source Deradji reste avec le sport **Translation** Deradji continue taking care of sport **Target**: يا سي دراجي انت مختص في الرياضة ابقا في الرياضة نبقو نحبوك **Translation** Mr derradji you are expert in Sport, please continue in sport and we will continue appreciating you
Source radja meziane vraiment cette chanson djat thebel be la voix dialek w rabi yerhem kamel messaoudi **Translation** Raja Meziane this song is wonderful with your voice, may God bless the soul of Kamel Messaoudi **Target** كلمات روعة وجات مع صوتك ربي يرحمو كمال مسعودي **Translation** Beautiful Lyrics especially with your voice, may God bless the soul of Kamel Messaoudi

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 for 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 4!
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

[1] A. Salama, H. Bouamor, B. Mohit, and K. Oflazer, "Youdacc: the youtube dialectal arabic comment corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, 2014, pp. 1246–1251.

[2] M. Abdul-Mageed, M. Korayem, and A. YoussefAgha, ""yes we can?": Subjectivity annotation and tagging for the health domain," in *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, 2011, pp. 666–671.

[3] O. Uryupina, B. Plank, A. Severyn, A. Rotondi, and A. Moschitti, "Sentube: A corpus for sentiment analysis on youtube social media," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, 2014, pp. 4244–4249.

[4] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How useful are your comments?: analyzing and predicting youtube comments and comment ratings," in *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, 2010, pp. 891–900.

[5] 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*, 2010, pp. 644–652.

[6] T. Etchegoyhen and A. Azpeitia, "Set-theoretic alignment for comparable corpora," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[7] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[8] D. Huang, L. Zhao, L. Li, and H. Yu, "Mining large-scale comparable corpora from chinese-english news collections," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10, 2010, pp. 472–480.

[9] I. Vulic and M. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, 2015, pp. 363–372.

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

[11] S. Abdul-Rauf and H. Schwenk, "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*, pp. 16–23, 2009.

[12] 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.

[13] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaïli, "Machine translation experiments on PADIC: A parallel arabic dialect corpus," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China*, 2015.

[14] S. U. Aqeel, S. M. Beitzel, E. C. Jensen, D. A. Grossman, and O. Frieder, "On the development of name search techniques for arabic," *JASIST*, vol. 57, no. 6, pp. 728–739, 2006.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[16] M. Saad, D. Langlois, and K. Smaïli, "Extracting comparable articles from wikipedia and measuring their comparabilities," *Procedia - Social and Behavioral Sciences*, vol. 95, pp. 40 – 47, 2013.