An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings

Karima Abidi, Kamel Smaïli

An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings

Karima Abidi, Kamel Smaïli
SMarT Group, LORIA, F-54600, France
{karima.abidi, kamel.smaili}@loria.fr

Abstract

The goal of this work consists in building automatically from a social network (Youtube) an Algerian dialect lexicon. Each entry of this lexicon is composed by a word, written in Arabic script (modern standard Arabic or dialect) or Latin script (Arabizi, French or English). To each word, several transliterations are proposed, written in a script different from the one used for the word itself. To do that, we harvested and aligned an Algerian dialect corpus by using an iterative method based on multlingual word embeddings representation. The multlinguality in the corpus is due to the fact that Algerian people use several languages to post comments in social networks: Modern Standard Arabic (MSA), Algerian dialect, French and sometimes English. In addition, the users of social networks write freely without any regard to the grammar of these languages. We tested the proposed method on a test lexicon, it leads to a score of 73% in terms of F-measure.

Keywords: Multilingual word embeddings, Algerian dialect, CBOW, comparability

1. Introduction

The wide use of social networks arises several new NLP issues: stretched letters, misspelled words, use of emoticons, condensed writing, etc. For the use of Arabic in social networks, the same phenomena are observed, henceforth other issues are also noticed and especially for Arabic dialects.

In this work, we are interested by the Algerian Arabic dialect. One needs to know that in Algeria, people speak their mother tongue, which is an Arabic dialect, but could speak the official language that is Modern Standard Arabic (MSA), French and sometimes English. Nevertheless, what makes the issue more challenging is that people can mix in the same sentence the whole previous languages.

Since this vernacular language is not written and no standardization exists, people write their comments in a free way. They do not pay special attention to grammar, consequently for uneducated people, they write a word as they want or, at the best, such as it is pronounced. People write sometimes Algerian dialect by using the Latin script (LS) because they are influenced by the French culture, this phenomenon is named in the community working on dialect: *Arabizi*. This could be explained also by the fact that in Algeria the mobile phone keyboards are in French that makes writing in Arabic more difficult. This obviously constitutes a serious issue because Arabic NLP tools could not be used on Latin script and more especially because the corresponding resources are not available.

Unfortunately, the NLP tools in French cannot be used either since the written words in LS are not necessarily in French. In the following, we give an example of a post extracted from YouTube written in LS:

{*ouiii 7na haka kima galha bn chnt w li maybghounach w ha l7itan ha l7itan*}.

That should be written in Arabic script (AS) such as:

{وي حنا هكا كيما قلها بن شنات و لى ميبغيوناش و ها لحيطان ها لحيطان}

that means: *We are like this such as Benchenet said: those who do not like us it does not matter it does not matter*.

Sometimes, people when they write in Latin script, they use some codifications for specific Arabic letters that do not exist in Latin. This is the case of ع that is replaced by *3*, ق that is written *9*, خ as *5* and other codes that are not officially adopted by everyone.

Another well known and frequent phenomenon in the Algerian dialect is the code-switching (Yoder et al., 2017)(Abidi and Smaïli, 2017a) that exist for other langauges (Dey and Fung, 2014). In Algeria people switch from the local Arabic to French or sometimes to MSA to express an idea in a well structured language. Switching may concern one isolated word or several contiguous words. For example, in our corpus we have found the following examples, which should be read from right to left:

{*les ingrédients* راهم وين يهديكم الله}

{*Algérienne qui aime le Maroc* وجعني قلبي الجزائر و المغرب بلد واحد}

In these two examples the writers started the sentences in Arabic and finished them in French.

In this paper, we are interested by the creation of a lexical resource containing for each entry, the corresponding different ways to write the same word. As presented before, a dialect word could be written in Arabic and Latin script. And for both of them, a word has different graphies since people write freely. For instance, in our corpus we found seven forms of the word بلادي:

{بلادي: بلدي, baldi, beladi, bledi, bladi, baladi }.

We propose in the following to use the concept of word embeddings in order to build a lexicon for Algerian spoken language, containing for each entry its different forms of writing. The entry could be a word in Algerian dialect expressed in Arabic script (dialect or MSA) or in Latin Script (Arabizi, French and sometimes in English). These forms are extracted from a large corpus harvested from YouTube.

The rest of the paper is organized as follows: Section 2. concerns the related work while Section 3. describes the collected corpus. In Section 4., we discuss the automatic method used to learn an Algerian Dialect Lexicon (ADL). In Section 5., we present a protocol to evaluate the rele-

vance of the extracted lexicon. Finally, in Section 6., we conclude and we discuss the future work.

2. Related work

The NLP community, which started few years ago to pay attention to Arabic dialects, is faced to the lack of resources. To remedy to this problem, the researchers often created them from scratch. Creating automatically a lexicon for Arabic dialects is then a challenging and important task for processing dialects. In (Al-Sabbagh and Girju, 2010), the authors propose to induce an Egyptian dialect lexicon by mining the Web. The idea is to create a lexicon of Egyptian dialects with their corresponding MSA synonyms. The approach used is based on retrieving collocation words in a large corpus. This approach leads to a lexicon of 1000 entries. This work is different from ours, but constitutes also an attempt to build automatically a lexicon. To the best of our knowledge, this resource does not exist for Algerian dialect, while it is necessary for different NLP tasks. In fact, for dialect identification, several methods based on machine learning have been used such as in (Belgacem and Zrigui, 2010) (Al-Badrashiny et al., 2015) (Harrat et al., 2015), but also a dictionary-based method could be considered, that is why a lexicon dialect may help to detect the origin of the Arabic dialect.

For Automatic Speech Recognition (ASR) of a dialect, a vocabulary is necessary to recognize the uttered sentences. In (Menacer et al., 2017), the authors adapted a MSA speech recognition system, but with French acoustic data. In fact, they have not succeeded to find Algerian data to adapt their ASR. The choice of using French data is motivated by the fact that Algerian dialect is highly code-switched as explained in the introduction. The same phenomenon is observed also in machine translation. In fact, in (Meftouh et al., 2015), the authors presented experiments on machine translation on PADIC (Parallel Arabic DIalect Corpus). Among them, experiments have been conducted on Algerian dialect. In this work, the authors have been confronted to a problem due to the fact that their lexicon has been induced from PADIC. The number of entries is weak and more especially a word in PADIC has only one way to write it, since the rules used to write PADIC were inspired from the way of writing MSA. This lack of varieties of a word leads to weak results in terms of BLEU. In fact, the training corpus was built by hand and not harvested from social networks. In (Harrat et al., 2014), the authors proposed several resources: morphological analyzer, diacritization and also an Algerian dictionary. This latter is composed of words extracted from the dictionary of BAMA (Buckwalter, 2002) that have been adapted to the dialect, but unfortunately the authors have not treated the Arabizi nor the foreign words.

The resource, we propose to develop, will help to solve one of the present phenomenon in natural language processing related to Arabic dialect: the profusion of texts written in Arabizi in social networks. This issue has been addressed in several works such as in: (Darwish, 2013), the author proposed an approach to identify and to convert Arabizi into Arabic characters. He used words and sequence-level features to identify Arabizi that is mixed with English. In (Al-Badrashiny et al., 2014), the authors presented a system that uses a finite state transducer trained at the character level to generate all the possible transliterations for the input Arabizi words. In (van der wees et al., 2016), authors proposed an Arabizi-to-Arabic transliteration pipeline that combines character level mapping with contextual disambiguation of Arabic candidate words.

In our work, comparatively to the three last ones, we consider a word written in Arabizi such as any other word in the Algerian dialect. Consequently, we do not want to identify it and make a particular treatment to convert it into Arabic script. In fact, the Arabizi represents the real world in the social networks and particularly for those used by the Maghrebi people.

3. Corpus

To build an Algerian dialect lexicon, we decided first to collect a large corpus from comments posted by people related to Algerian videos. That is why, we harvested data from YouTube by using the Google's API [1] that allows users to search for videos that match specific criteria and retrieve all information and comments of these videos. To harvest, we chose few keywords to form queries in order to retrieve videos concerning national news, Algerian celebrity, local football, etc. Table 3. shows some figures before and after preprocessing the collected data, where $|C|$ is the number of comments, $|W|$ is the number of words and $|V|$ is the vocabulary size.

	Raw corpus	Cleaned Corpus
$\|C\|$	1.3M	1.1M
$\|W\|$	20M	17.7M
$\|V\|$	1.3M	0.99M

Table 1: The collected YouTube Algerian Dialect Corpus.

We can mention that after the cleaning process, the corpus has been reduced by around 15% and the vocabulary by around 24%.

4. Lexicon learning method

Because people, in social networks, do not mind about the spelling, a word may be written according to its pronunciation or to the one supposed by the user. One can borrow French words with foreign letters correponding to sounds the users have not in his tongue-mother such as /p/, /v/ and /g/ and adapt them to the dialect. For automatic speech recognition this could constitute a problem since the original pronunciation of the word is altered. For example, the French word *Problème* (*/problɛm/*) will be pronounced in Algerian dialect */broblɛm/*. Transliteration of foreign words are let to the goodwill of the users. Our objective is to produce automatically all the different forms of a word according to the writing varieties presented in the introduction. Each entry of this lexicon will be associated to all its different forms of writing harvested from YouTube. Another

[1] Available at: https://developers.google.com/YouTube

important motivation for learning automatically this lexicon is the fact that in dialect, people create new words frequently. That is why learning automatically such a lexicon is necessary to cope with the dynamic of the evolution of the lexicon.

4.1. On the need of comparable dialect corpus

In order to identify words, which are related to each others, we need to build automatically a comparable corpus. In a previous work (Abidi et al., 2017), we addressed the difficult issue of matching comments from YouTube for a vernacular language (Algerian dialect) for which no writing rules do exist. The method we propose is based on the concept of learning multilingual word embeddings (Word2Vec). The objective is to find a list of words that could be correlated to a lexical entry whatever the language. This method has permitted to find a list of variations of the same word. Then these words have been exploited in the matching process of documents. The word2Vec approach has been iterated to improve, at each step, the quality of the supposed comparable documents. This method achieved good results and allowed us to build a comparable Algerian dialect corpus named CALYOU: A comparable spoken Algerian corpus extracted from YouTube.

4.2. Training method

In the following, we propose to detail the method we developed to generate an Algerian dialect lexicon. Because the lexical variability in Algerian dialect is very high, in other words, each word could be written in several ways and because, the dialect evolves frequently, we propose to learn the dictionary automatically from social networks. Regularly, this dictionary could be enhanced by running again the proposed method.

The collected corpus from YouTube is transformed into a comparable document by gathering the comments, which are similar by using the method proposed in (Abidi et al., 2017). In fact, the comparable corpus is obtained in an iterative process where at each step, we refine the quality of the comparability of the corpus. At each iteration, two vocabularies are produced: a Correlated Word Lexicon (CWL) and an Algerian Dialect Lexicon (ADL) (see Figure1). CWL is used to produce comparable documents and ADL is the expected Algerian Dialect Lexicon.

CWL and ADL are modified at each iteration, when CWL is refined, the quality of the comparable corpus is improved and a fortiori, the entries of ADL will be more and more precise. Each entry of ADL will be represented by a word and its different ways of writing it. An entry in this dictionary could be written in Arabic (MSA or dialect) or in Latin script (Arabizi, French and sometimes in English).

4.2.1. Learning CWL

For learning CWL, we decided to use Word2Vec to retrieve the correlated words. To do that, for each word (w_s) of the corpus, where s is the Arabic or the Latin script, we learned its correlated words ($w_{\bar{s}}$), where $\bar{s}$ is a script different from s. We opted for a continuous bag of words (CBOW) method (Mikolov et al., 2013) with a sliding window of 100. This size has been fixed after several tests. This large number is explained by the fact that all the comments concerning the same video have been concatenated into one document (Abidi and Smaïli, 2017a).

For each w_s, we keep its n best correlated words $w_{\bar{s}}$. This process is used for each word of the corpus, at the end, we achieve a list of words and their n best correlated words. From this list, all the entries that occur more than α are inserted into CWL. For the others, for each w_s^i, all the words $w_{\bar{s}}^j$ respecting the following constraints:

$$\begin{cases} N(w_{\bar{s}}^j) > \alpha \\ or \\ S(w_s^i) = S(w_{\bar{s}}^j) \end{cases} \quad (1)$$

are included into CWL. Where $N(x)$ is the occurrence of x, α is set empirically to 1000 and S is a function that encodes phonetically a word. The constraint represented by S takes into account the variability of writing a form in accordance to its pronunciation. For that, the function S is implemented by using Soundex (Aqeel et al., 2006), a phonetic algorithm for indexing by sound. Words are encoded by taking advantage of their phonetic form. If two words have the same code, we can conclude that, one is the transliteration of the other. Soundex proposes to 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. We obviously adapted the original correspondence table in order to take into account the graphemes of our corpus (see Table 2). The characters of

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

Table 2: Encoding correspondence table

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, the encoding process of the dialect word **حومة** in accordance to the encoding Table 2 and to the transliteration Table 3, will propose two codes: *H500* and *7500*. While the words *Houma* and *7ouma* will be encoded respectively *H500* and *7500*. Consequently, this allows to make a correspondence between the three latter words: **حومة**, *Houma* and *7ouma*. For more details see (Abidi and Smaili, 2017b). In Table 4, we give an example of CWL entry [2].

4.2.2. Learning ADL

As for training CWL, we used the CBOW (Mikolov et al., 2013) approach to constitute the Algerian dialect lexicon

[2]The translation in English are proposed by the authors for readability

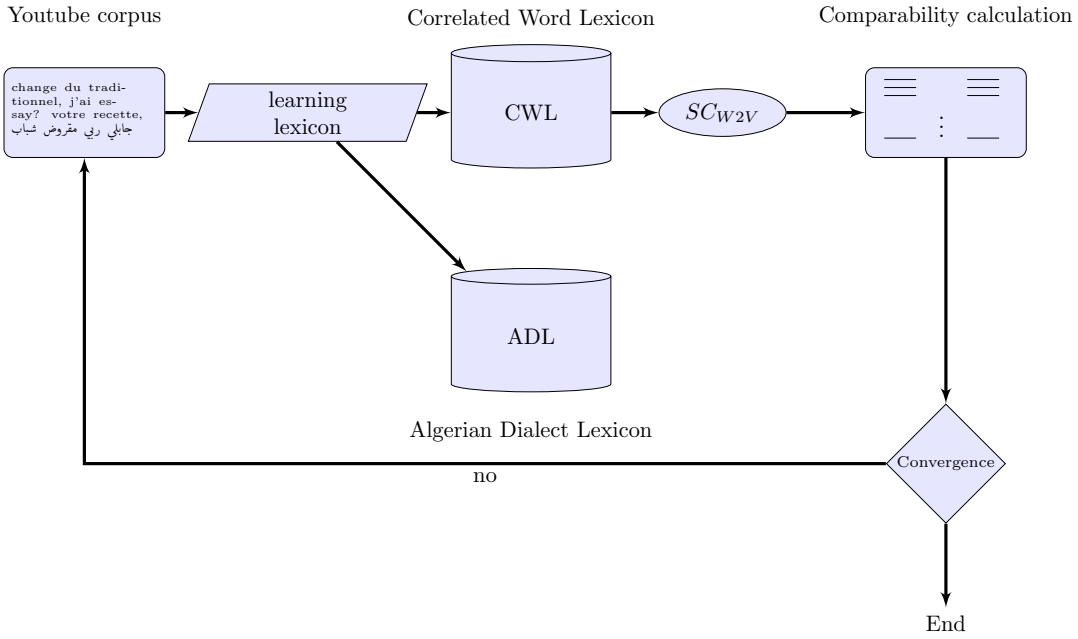

Figure 1: The training process

Arabic letter	Trans	Arabic letter	Trans	Arabic letter	Trans
ا	2 a e i	س	s	ع	‘ 3 a e
ب	b p	ش	sh ch	غ	r gh
ت	t	ص	s	ي	y i
ث	t th	ض	d	ذ	d dh
ج	j dj	ط	6 t	ر	r
و	w u ou	ك	k c	ق	9 q k c
خ	5 kh	ل	l	ظ	d
ح	7 h	م	m	أ	a
د	d	ن	n	أُ	u o
ف	f v	ه	h	إ	i
ز	z	ة	h a	ا	a

Table 3: Codes of Arabic-Latin transliteration

سعود (*Saud*)	saoud (*Saoud*) therories (*terrorist*) saudi (*Saudi*) alzarooni(*Al_Zarouni*) saoudi(*Saudi*) iran (*Iran*) terrorist (*terrorist*)

Table 4: An example of CWL

(ADL). In order to do so, a word (w_s) and its n best correlated words ($w^j_{s+\bar{s}}$) are retained for the next treatments. Where $s + \bar{s}$ means any kind of script (Arabic or Latin). This process is used for each word of the corpus, which achieves a list of words and their n best correlated words. n is determined empirically and in our experiments it has been fixed to 40. From this list, each entry will be processed to find its accurate declensions and will constitute, if appropriate, an ADL entry. Two cases have to be examined.

Same script of words

An entry w^i_s with its correlated words w^{ij}_s, which are written in the same script are included into ADL, if they respect the following constraint: $R(w^i_s) = R(w^{ij}_s)$. Where $R(x)$ is a function which removes vowels from x. This is motivated by the fact of the high ambiguity of writing in Arabizi since we write in Arabic by using another alphabet. In fact, a user who writes in Arabizi, sometimes does not find easily the exact Latin sound corresponding to the Arabic one, consequently in replacement he allows himself to take what he considers being the closest sound in Latin script. This operation will help to capture a word and its different ways to write it.

Different script of words

A new entry is inserted into ADL, composed of w_s and $w_{\bar{s}}$ with s corresponding to the Latin script if they respect the following constraint: $\exists i, w^i_s \in L(T(w_s))\ and\ w^i_s = w_{\bar{s}}$. With $T(x)$ is the transliteration of x and $L(T(x))$ is the list of possible transliterations of x. The transliteration is done in accordance to Table 3. This allows to associate a word written in Latin with words written in Arabic script .
The use of the procedures mentioned before allows to produce an Algerian Dialect Lexicon. An example is given in Table 5. One can remarks that, in comparison to Table 4, the entry of the word سعود *(Saud)* in ADL is more accurate. Only words related to this entry are kept and other words that are correlated such as *terrorist* and *Iran* are discarded.

سعود (Saud)	sa3oudi (Saudi) saud (Saud) saudi (Saudi) saoudia(Saudi Arabia) saoudi(Saudi) soudia (Saudi Arabia) saudia (Saudi Arabia) saoud(Saud)

Table 5: An example of ADL

5. Results

It is difficult to evaluate automatically the quality of a lexicon produced by an automatic method. Even if we do

have a measure, we need a reference test lexicon to evaluate it. For our experiments, we decided to test the quality of the produced lexicon by using, the classical retrieval information: Recall, Precision and F-measure, on a test corpus. Furthermore, we tested also the evolution of the number of words in the lexicon, in accordance to the iterative Word2Vec process.

To calculate the F-measure, we need a reference test lexicon. To the best of our knowledge, there is no Algerian dialect lexicon similar to ADL, on which we can evaluate our method. That is why, we decided to build semi-automatically a reference test lexicon. We collected data from social networks, and for each word, we used the algorithm based on Soundex described in Section 4.2.1. to get all the words sharing the same phonetic codes. Then this list is cleaned and updated by a human being. To illustrate this, in Table 6, we give an example concerning the word يشفيها (*It cures her*) obtained by Soundex. In this list, some words are similar to the entry, but others (those which are written in bold) have different meanings that should be discarded. Consequently, a cleaning process is done on each entry: removing the words (bold examples in Table 6) that are not related to the entry, removing the words (in italic in Table 6) which are similar, but they do not have the exact meaning as the entry and adding the missed words (examples in blue in Table 7).

This process led to a test lexicon of 560 entries, with an average number of forms by entry of 6, a maximum of 17 and a minimum of 1.

يشفيها	*ychafi ychafih* **yechf ychoufo yechfo** yachfih ichfiha ychafiha yachfiha *ychafih* ichafiha *yachefih*

Table 6: An example of an automatic extraction of a potential entry of the test lexicon

يشفيها	ychafiha yachfiha ichafiha ichfiha yechfiha ychfiha yechafiha

Table 7: An example of an entry of the test reference lexicon

In Figure 2, we plot the evolution of Recall, Precision and F-measure values for each Word2Vec iteration. The curves show clearly that the three measures progress. From the 17^{th} iteration, the three values are close to each others and from the 20^{th} iteration, the Precision decreases. Since we would like to propose an accurate lexicon, we decided to stop the iterative process, when the Precision starts decreasing. At the beginning of the training process, the F-measure is bad, since at the initial step, the CBOW method runs on articles composed of bulk comments. Theses comments are dispatched over the documents, consequently the CBOW process is not able to retrieve similar words in terms of graphemes, which makes the retrieved Algerian dialect lexicon not accurate. When we align the comments and by injecting them into the learning process, we get a better training corpus that leads to an Algerian dialect lexicon with more entries as illustrated in Figure 3. This curve illustrates the evolution of the number of entries in the Algerian Dialect Lexicon for which 85% of the entries have been added in the ten first Word2Vec iterations. But as mentioned before, the iterative process should be stopped when the Precision starts decreasing.

In Figure 4, we illustrate the added number of forms be-

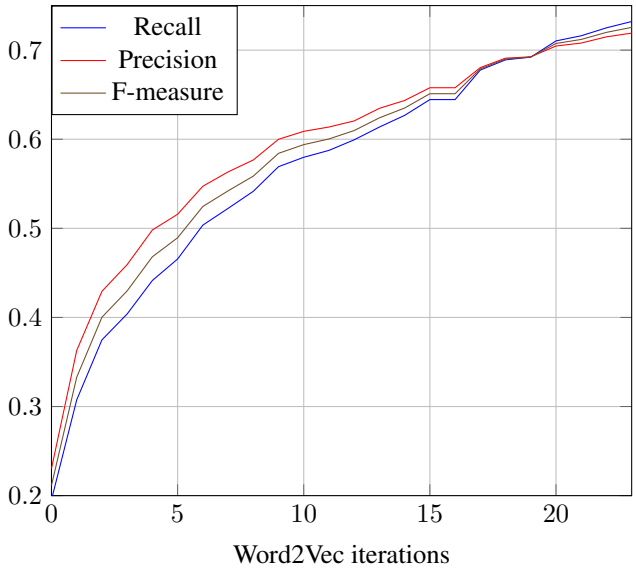

Figure 2: Evolution of the Recall, Precision and F-measure for each Word2Vec iterations

tween the first and the last launch of the iterative Word2Vec process. One remarks that, when Word2Vec is launched at the beginning (gray bars), 84 entries have 5 different ways to be written, while at the end of the process (black bar), this number increases to 320. The number of entries in the dictionary having more than 5 forms, at the first launch of Word2Vec process, is equal to 201. At the end of the process, this number jumps to 1145. This figure shows that the distribution of entries with only one transliteration, at the end of the iterative process, is 35%, while the distribution of entries with more than 30 forms represents 7%. This last result strengthens the fact, that the variability of the Alge-

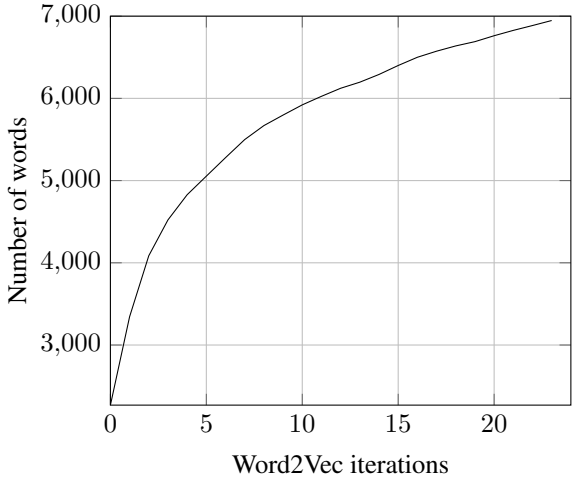

Figure 3: The evolution of the number of words in ADL in terms of Word2Vec interations

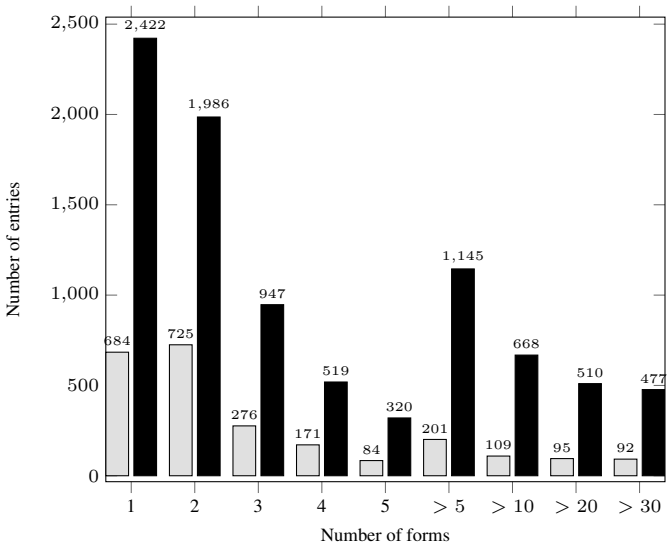

Figure 4: The progression of the number of forms between the first and the final iterative Word2Vec process

rian dialect is very high and consequently it necessitates an automatic process to build a dialectal lexicon.

In figure 5, a part of the built ADL is illustrated. One can remark that the entries are written either in Arabic or in Latin script. For each entry there is one or several forms (transliterations). Some entries are in French, but some of them are miswritten as explained in this paper, such as *comentair* which should be written *commentaire*. The entry مستر corresponds to the English word *mister*, one can notice that some transliterations are not well written in English, but this corresponds to how the users wrote them. Some entries have several transliterations such as يرحمك, which has 67 different forms.

```
<?xml version="1.0" encoding="UTF-8"?>
<Lexique Nb_Translitteration_possibles="6947">
<"شوربة" Transliteration=" chorba" />
<"مستر" Transliteration=" mester mister mstr" />
<"ميستر" Transliteration=" mester mister mstr" />
<"يرحمك" Transliteration=" yr7mk yr7mak yrhamak
 yarhamek yarhemak irahmk yarhamk yr7mek yere7mek yarhamak
 yarhemek yrhmk yar7mak yarhmak yarhmek yar7mik
 yarahmak yar7mek yarahmk yerhemek yarahmek yerehmek yerhamek
 yer7mak yare7mek yerhamak yer7mek yerehemek yarhmeke rahimaka
 yrahmek yrahmak irahmak irhmak irahmek yra7mk irhmk yrehmak
 yera7mak yerehmk yera7mek yrehmek yara7mak yarehmek yara7mek
 yerahmeke yarehmak yarhmk yerhmk yarhmeek yra7mak ir7mak
 yra7mek yarhamoka yrehmk yar7mk yrahmk ira7mak irehmek
 yerhmek yerahmk yrhmek yerahmek yerhmak yerahmak yrhmak
 yarahemek" />
 <"كاشير" Transliteration=" kachir" />
 <"film" Transliteration="فيلم فلم فليم" />
 <"comentair" Transliteration="كومونتار كومنتار" />
```

Figure 5: Example of some entries of the ADL produced by the iterative Word2Vec

6. Conclusion

In this article, we present an iterative multilingual word embeddings approach, which allowed to make comparable an Algerian dialect corpus, from which we built automatically a lexicon. Each word of this lexicon is associated to its different transliterations, the method led to a dictionary of 6947 entries. An entry may have a minimum of one transliteration and a maximum of 71. We observed, that 7% of the entries have more than 30 transliterations. This dictionary has been tested by using the Recall and Precision measures on a lexicon of 560 entries built semi-automatically. The iterative method of building the dictionary has been stopped when the precision has started decreasing. This method achieved a F-measure of 73%. Since the dialect is evolving everyday, one of the advantage of this approach is that the lexicon can be updated easily by harvesting new data. Also, this method could be used for any dialect for which data are available in the corresponding social network.

In the best of our knowledge, this kind of dictionary does not exist, it will be useful for different applications, for instance identifying parallel segments in comparable documents. It could be used to develop a based-dictionary transliteration system for Algerian dialect. In machine translation, this resource might be used to establish a correlation between a word and its corresponding transliterations and especially for proper names. This resource (ADL) is freely available at `http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/Lexicon`.

7. Bibliographical References

Abidi, K. and Smaïli, K. (2017a). An empirical study of the Algerian dialect of Social network. In *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*, Casablanca, Morocco.

Abidi, K. and Smaili, K. (2017b). How to match bilingual tweets? *Sixth International Conference on Natural Language Processing 25-26, 2017 in Sydney, Australia*.

Abidi, K., Menacer, M. A., and Smaïli, K. (2017). Calyou: A comparable spoken algerian corpus extracted from youtube. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm Sweden August 20-24 2017*.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 30–38.

Al-Badrashiny, M., Elfardy, H., and Diab, M. T. (2015). AIDA2: A hybrid approach for token and sentence level dialect identification in arabic. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–51.

Al-Sabbagh, R. and Girju, R. (2010). Mining the web for the induction of a dialectical arabic lexicon. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Aqeel, S. U., Beitzel, S. M., Jensen, E. C., Grossman, D. A.,

and Frieder, O. (2006). On the development of name search techniques for arabic. *JASIST*, 57(6):728–739.

Belgacem, M. and Zrigui, M. (2010). Automatic identification system of arabic dialects. In *Proceedings of the 2010 International Conference on Image Processing, Computer Vision, & Pattern Recognition, IPCV 2010, July 12-15, 2010, Las Vegas, Nevada, USA, 2 Volumes*, pages 740–749.

Buckwalter, T. (2002). Buckwalter arabic morphological analyzer version 1.0. *Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.LDC2002L49*.

Darwish, K. (2013). Arabizi detection and conversion to arabic. *Computing Research Repository*, abs/1306.6755.

Dey, A. and Fung, P. (2014). A hindi-english code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 2410–2413.

Harrat, S., Meftouh, K., Abbas, M., and Smaïli, K. (2014). Building Resources for Algerian Arabic Dialects. In *15^{th} Annual Conference of the International Communication Association Interspeech*, Singapour, Singapore. ISCA.

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., and Smaïli, K. (2015). Cross-dialectal arabic processing. In *Computational Linguistics and Intelligent Text Processing - 16^{th} International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I*, pages 620–632.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaïli, K. (2015). Machine translation experiments on PADIC: A parallel arabic dialect corpus. In *Proceedings of the 29^{th} Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*.

Menacer, M., Mella, O., Fohr, D., Jouvet, D., Langlois, D., and Smaili, K. (2017). Development of the arabic loria automatic speech recognition system (alasr) and its evaluation for algerian dialect. In *Third International Conference On Arabic Computational Linguistics, Dubai, November 2017*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

van der wees, M., Bisazza, A., and christof Monz. (2016). A simple but effective approach to improve arabizi-to-english statistical machine translation. page 43.

Yoder, M. M., Rijhwani, S., Rosé, C. P., and Levin, L. (2017). Code-switching as a social act: The case of arabic wikipedia talk pages. In *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 73–82, Vancouver, Canada, August 3, 2017.*