ALG/FR: A step by step construction of a lexicon between Algerian Dialect and French

Faical AZOUAOU
Ecole Supérieure d'informatique d'Alger ESI

f_azouaou@esi.dz

Imane GUELLIL
Ecole Supérieure d'informatique d'Alger ESI
Ecole préparatoire des sciences et techniques d'Alger EPSTA

i_guellil@esi.dz

Abstract

Arabic is the official language spoken by 350 million people living in 22 countries in the world. However, this language is in state of diglossia where it lives with many dialects. A lot of research works have been done on the treatment of Arabic dialects. The Arabic Dialects are considered as under-resourced languages. Hence, many works are orientated to the resources construction (So construction of lexicon and corpora).

Algerian Dialect is a Maghrebi Dialect spoken by over 40 million people. This dialect is rich and complex. It contains Arabic words, French words, Turkish words,…etc. However, it has only few works that treated it. Concerning resources construction, we got only one parallel corpus, one monolingual corpus and one lexicon between this dialect and Modern Standard Arabic (MSA). However this lexicon is not a freely resource, so we could not exploit it.

In this paper, we construct an Algerian/ French lexicon. This lexicon contains the translation of general words and sentiment words from Algerian Dialect to French. We begin with 1144 words and apply some orthographic rules concerning Algerian dialect to enrich the initial lexicon and finally obtain 25 086 words.

1 Introduction

The Arabic language is rich and complex. Its treatment represents a significant challenge for the NLP. Arabic is the official language in 22 countries. It is spoken by over 350 million people over the world. However the Arab language is in a state of diglossia in these countries where Standard Arabic (MSA) and regional dialects are closely related. So, the Arabic language has 22 dialects (Sadat et al. 2014A). These dialects are grouped into six categories: EGYPTIAN (for Egypt) LEVANTINE (for Syria and Palestine), GULF (for golf country), IRAQI (for Iraq) and MEGHREBI (for North Africa countries) and others (Zaidan and Callison-Burch 2014).

Recently considerable interest has been given to the treatment of Arabic dialects. This is mainly due to their wide use in social media (Sadat et al. 2014A), (Al-Sabbagh and Girju 2012b), (Shoufan and Al-Ameri 2015). Unlike (MSA), Arab dialects have not a predefined set of grammatical rules. These dialects are themselves very distant from each other. For example, the Moroccan dialect (Maghrebi) is difficult to understand by LEVANTINE speakers (Zaidan and Callison-Burch 2014).

Several categories of work were conducted for the treatment of these dialects begging by basic analysis (morphological analysis, syntactic ...) (Habash and Rambow 2007) (Sadat et al. 2014b), to achieve semantic analysis (automatic translation and sentiment analysis of social media users) (Jehl et al. 2012), (Abdul-Mageed et al. 2014). For our part, we first concentrate on resource construction (specifically lexicon). We believe that before starting any work on a given language or dialect, we must first have the resources. In addition to this, some dialects are considered without resources, this can be illustrated by the paper of Meftouh et al. in (. Meftouh et al 2012), titled "A study of a non-resourced language: an Algerian dialect." In this paper the authors clearly affirm that the Algerian dialect is a dialect without resources.

Some work have however conducted on resource construction (lexicon and corpus) from Arabic dialects. For example, the construction of lexicon was done for the Iraqi dialect in (Graff et al. 2006), for Tunisia dialect in (Boujelbane et al. 2013). For the corpus construction, we found that most of the work manually develop annotated corpus (Al-Sabbagh and Girju 2012b) and (Al-Sabbagh and Girju 2012a) who focus on an Egyptian corpus or (Zaidan and Callison-Burch 2014) where the authors make use of several annotators that speak sever Arabic dialect to label 330,930 sentences of several dialects.

We have however been able to identify a minority of work on the Algerian dialect. To respond of all the cited problems, we propose ALG/FR, a bilingual lexicon between an Arabic dialect and other languages. This lexicon will serve us later to dialect identification task and for automatic translation of this dialect to another language or to analyze the users sentiments concerning a given product on social media. For responding to the lack of work on the Maghrebi dialects, we take as application case the Algerian dialect. Since, Zaidan et al. In (Zaidan and Callison-Burch 2014), affirm that the Maghrebi is close to the French unlike other dialects that were close to the MSA, our lexicon contains the French as a second language translation. Based on our survey in (Guellil and Boukhalfa 2015) that expose some problematics that have been treated before begin research on opinion mining and sentiment analysis, we first concentrate on Arabic dialects treatment. This treatment will be used later as part of our work on opinion mining and sentiment analysis in the social media.

2 Related work

During the last ten years, the interest of Arabic dialects treatment has greatly increased. This increase is mainly attributed to the large use of these dialects in social media (Shoufan and Al-Ameri 2015). The survey of Shoufan et al (Shoufan and Al-Ameri 2015) presents the work that treated these dialects by grouping them into four categories: basic Analysis, resource construction, identification of Arabic dialects and semantic analysis on these dialects.

By basic analysis, the authors of this survey targeted work addressing the morphological, syntactic and orthographic analysis (Guellil and Azouaou 2017a). The construction of resources concerns either lexicon (monolingual, bilingual or multilingual) or corpus (comparable or parallel corpus). For the identification of Arabic dialects, this survey focuses on two types of identification: textual (Guellil and Azouaou 2015) and vocal. These authors conclude semantic analysis work. These work are dedicated to machine translation (Guellil and Azouaou 2017b), sentiment analysis and subjectivity analysis of messages written in Arabic dialects.

As part of our work, we first based on the survey of (Shoufan and Al-Ameri 2015). Our primary goal in this paper is the construction and enrichment of a resource (bilingual lexicon) for processing Arabic dialects. So we focus in this part on the work related to construction resources (lexicon and corpus). We classify this work from the five families of dialects treated by the research community, that is: the Egyptian (EGY, for Egypt), the Levantine (LEV concerning Jordan, Lebanon, Palestine and Series), the dialect of golf country (GULF concerning el-Bahrain, Kuwait, El-Qatar, Oman, ...), the maghrebi (MAGHEREBI for Algeria, Tunisia, Morocco, Libya and Mauritania) and the Iraqi (IRAQI for Iraque). We end our analysis with the work focusing on Algerian dialect.

However, before presenting all these work, we first illustrated in Figure. 1, the representation of different dialects related to their country. This figure was taken from work of Zaidan et al. in (Zaidan and Callison-Burch 2014).

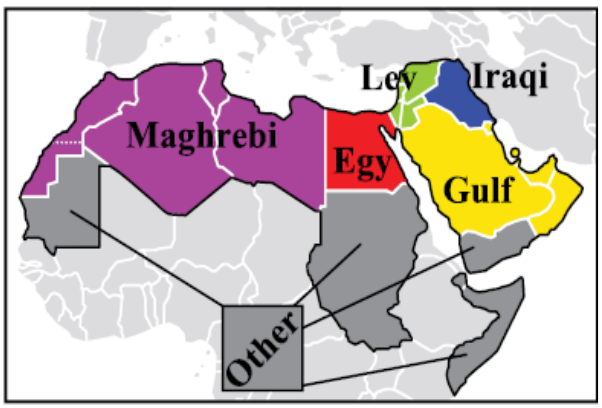

Figure 1: Different categories of Arabic Dialect

The problem of lack of resources for Arabic dialects treatment is really known by the research community. Several studies attempt to address this problem by using constructing resources: lexicons or corpus (Shoufan and Al-Ameri 2015). Zaghouani et al. in (Zaghouani 2014) presents a

survey on different resources treated MSA and Arabic dialects (lexicon and corpus) available in the net. To our part, we rely on these two types of resources (lexicon and corpus) and present the various work concentrating on their construction.

2.1 The work on lexicon construction

Within this category of work, we focus on building lexicons and dictionaries. Note that we classify work compared to their Arabic dialects category.

For EGY, we cite work of Diab et al. in (Diab et al. 2014), focusing on the lexicon construction named **Tharwa**. This lexicon is between the MSA and the Egyptian dialect. It provides deep and rich linguistic information on each of these entries, so: POS, ROOT, Pattern, and its translation into English and MSA. This lexicon was constructed based on several existing resources, such as: EGY-English dictionary, dialect EGY dictionary, the MSA-English dictionary and a morphological analysis of the Egyptian system. The resulting lexicon contains 73,348 words.

Regarding the LEV, we cite the work of Duh et al. in (Duh and Kirchhoff 2006) concerning the lexicon construction for the treatment of LEV dialect. This construction starts with the analysis of LEV words comparing the MSA, and with a LEV analyzer. However, the authors found that the LEV differs significantly from the MSA.

However, we observe that for Maghrebi, practically all paper concerns the Tunisian dialect (TUN). We start by paper of Boujelbane et al. in (Boujelbane et al. 2013). The authors of this paper are focused on creating resources (lexicon and corpus) for the translation of Tunisian dialect. In the lexicon part, they describe a methodology for the creation of a bilingual dictionary between the Tunisian dialect and MSA. However the authors of this paper are focused their eforts on the verbal part of the lexicon. The second work that we cite (Hamdi et al. 2015) concerning the creation of a POS tagger for a given language using the resources of another. The authors concentrate on Tunisian Dialect too. Due to the lexical difference between the MSA and the Tunisian, the conversion process is limited. To address this limitation, the authors construct three types of lexicons: 1) Verbal lexicon, 2) lexicon of verbal nouns and 3) lexicon of particles. The verbal lexicon contains 29,911 verbs. The verbal nouns lexicon automatically built from the lexicon of verbs because we can build adjectives, nouns, infinitive forms, etc from the verbs. The lexicon nouns contain 33,271 words. The particle lexicon contains 262 words and concern conjunctions, prepositions, etc.

For IRAQI, we cite the work of Graff et al. in (Graff et al. 2006), focus on six different lexicons of the Iraqi dialect. The result of this work contains a complete set of pronunciation, morphology, POS and annotations in English. It contains 120,000 words.

We finish with give a particular attention to Algerian dialect and we cite the work of Harrat et al. in (Harrat et al. 2014). The purpose of this paper is to make a translation between the Algerian dialect and classical Arabic (MSA). This work focuses on two kinds of Algerian dialect: (ALGR) which is the dialect of Algiers and (ANB) which is the dialect of Annaba (a city of Algeria). These authors will therefore build two lexicons: the first between MSA-ALGR (with 10790 words) and the second one between the MSA-ANB (with 9688 words. The authors finish they work by affirming that it is more difficult to translate the ALGR to the MSA that the ANB to the MSA. This is because the ANB is closer to the MSA that ALGR (the authors clearly shown that by using the Levenshtein distance).

2.2 The work on corpus construction

Regarding EGY dialect, we start with the work in (Al-Sabbagh and Girju 2012b), focusing on construction of corpus for multi kind Arabic dialect. However, these authors focus on the Egyptian dialect. To achieve this corpus, authors initially extracted Egyptian dialect corpus of Twitter, newspapers and blogs. It also uses a function of dialect identification, which analyzes the different phonological changes to propose at the end a POS tagger. Within the same dialect category, we can also mention the work Bouamor et al. in (Bouamor et al. 2014), building the first multi corpus dialect. This corpus contain 2000 sentences with their translation English, MSA, and the various dialects EGY, TUN (for MAGHEREBI) and Jordanian, Palestinian and Syrian (for LEV).

For the LEV, we present firstly, the work in (Bouamor et al. 2014), that propose a multi-dialect corpus also contains the LEV. We also cite (Maamouri et al. 2006) where the authors describe a methodological procedure for the development of a corpus between LEV/MSA. This corpus contains a morphological and syntactic annotation of approximately 26000 words.

For GULF dialect, we can cite the work in (Almeman and Lee 2013), who built a multi kind corpus of Arabic dialect using a corpus extracted from the web. The authors collected a total of

48,000 000 words, whose 145,000 000 for GULF dialect, 104,000,000 for the LEV dialect, 13 000 000 for EGY and 101 000 000 for the Maghrebi.

For Maghrebi, we cite again the work in (Boujelbene et al. 2013) that uses the lexicon built for the generation of corpus between the Tunisian dialect and MSA. This corpus will be used to enrich the lexicon that was built semi-automatically. We also cite the (Bouamor et al. 2014), concerning TUN dialect and others.

We finish with Algerian dialect and we mention two work (2014 Cotterell et al.) And (2012 Meftouh et al.). The first one of Cottrell et al. in (Cottrell et al. 2014), includes a new problem, which is Code-Switching (CS). CS represents a linguistic phenomenon where the speakers mix multiple languages within the same speech. The primary goal of the work (Cotterell et al. 2014) is to build a corpus between the Algerian dialect and French with integrating CS. For this, the authors firstly extract automatically 598,047 pages of Algerian newspapers in September 2012. Then, they proceed to a conversion of Arabic letters Arabizi (letter used in the Algerian dialect proposed (Yaghan 2008)). The provided corpus is the first large corpus for the Algerian dialect. The work of Meftouh et al. in (Meftouh et al. 2012) is certainly the first work dealing with the Algerian dialect. They focus on the dialect of Annaba ANB (Algeria EAST). The authors also present the methodology to build a parallel corpus between the MSA and the Arabic dialects. This work was subsequently improved in (Harrat et al. 2014), to provide a translation system between the MSA and the Algerian dialect.

In order to synthesize the studied work we propose Table 1. We classify in this table the various work related to construction resources (lexicon or corpus) from the each dialect categories (that we before present). We distinguish between the five families studied dialect (EGY, LEV, GULF, MAGHEREBI and IRAQI) while we focus on the Algerian dialect. By analyzing this table, we found that the only lexicon built for the Algerian dialect is done in (Harrat et al. 2014). However this lexicon was built between ANB-MSA and ALGR-MSA. The authors of this work affirm that the Algiers dialect (ALGR) is very far from the MSA and that is why their system works best with the ANB. Based on the work of Zaidan et al, (Zaidan and Callison-Burch 2014), we can say that the Algerian dialect is closer to the French unlike other dialect that closer to MSA. Based on the work Meftouh et al. in (Meftouh et al. 2012), we can say that the Algerian dialect is a dialect without resources (Lexicon and corpus) for processing it. For all these reasons, we propose in this paper the construction step by step of a bilingual lexicon ALG/FR used for Arabic dialect treatment. However, we focus on the Algerian dialect (which have a lack of resources) and include the translation of each term of this dialect in French (since this dialect is closer to the French). Contrary to other work on Algerian dialect treatment, this lexicon can be used in a social media.

	Construction resources	
	Lexicon	**Corpus**
EGY	(Diab et al. 2014)	(Al-Sabbagh and Girju 2012b) (Bouamor et al. 2014) (Almeman and Lee 2013)
LEV	(Duh and Kirchhoff 2006)	(Bouamor et al. 2014) (Almeman and Lee 2013)
GULF		(Almeman and Lee 2013)
MAGH -REBI	(Boujelbane et al. 2013) (Hamdi et al. 2015)	(Bouamor et al. 2014) (Almeman and Lee 2013)
IRAQI	(Graff et al. 2006)	
ALG	(Harrat et al. 2014)	(Meftouh et al. 2012) (Cotterell et al. 2014)

TABLE 1: Synthesis of the studied work

3 Contribution: construction and enrichment of a bilingual lexicon between the Algerian dialect and French

The main contribution that we propose in this paper is the construction and enrichment of a bilingual lexicon (between the Algerian dialect, mainly Algiers, and French). Note that, this work represents one component in our family of work concerning opinion mining and sentiment analysis. So, in our lexicon we add the sentiment words. To achieve this lexicon, we follow three steps. 1) Extraction and improving a dialect lexicon and extraction of the most intense word from a sentiment lexicon. 2) Merge the two lexicons (dialect end sentiment lexicon) and replace some letters with the most used letters in social media. 3)

Enrichment lexicon obtained using spelling variations that appear in social media. Figure. 2 show clearly the transition between these different steps. We show in the following each of these steps.

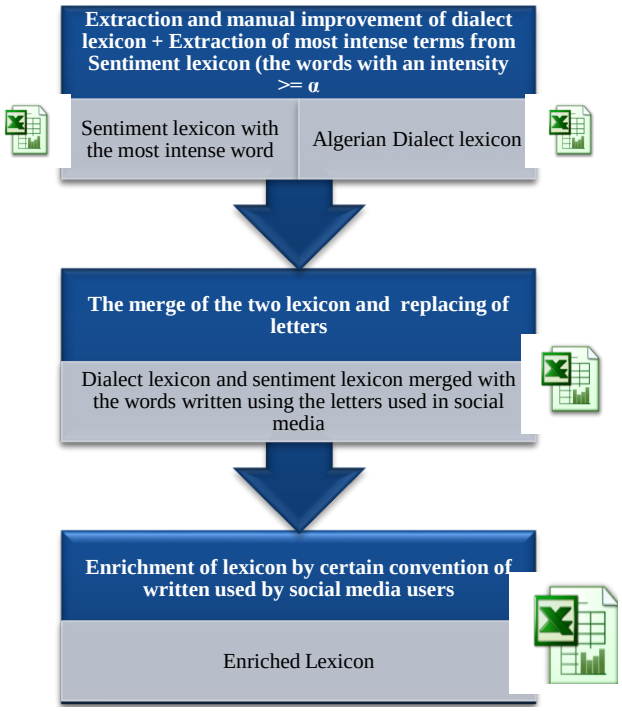

Figure 2: **construction and enrichment of a bilingual lexicon between the Algerian dialect and French**

3.1 Extraction and manual improvement of dialect lexicon and extraction of most intense terms from Sentiment lexicon

We first extract the only existing lexicon (to our knowledge) on the net between the Algerian dialect and French. It contains the translation of most used verbs, nouns and adjectives in Algerian dialect. We then manually analyzed a corpus of comments of an Algerian page from social media. The first observation we can make is that the users of this page, except verbs, adjectives and nouns, rely heavily on conjunctions, relative pronouns, etc. Our first contribution is the manual construction (of a particle list, so list of: conjunction, pronoun, etc). An example of a particle is: *fi* (in) *Fouk* (on) or *nchalah* (if god want, which is used a lot by the Algerians). Let us simply note that the generation of particles was also doing in the work of Hmidi et al. in (Hamdi et al. 2015) for the Tunisian dialect (TUN).

Remember that our work represents one brick of a family of work to analyze sentiment users from given product within social media. In this context, it is interesting to incorporate terms related to feelings or sentiments such as love, hate, great, sad, etc. Given that sentiment lexicon which are very large, we opt for an approach selecting words having a higher intensity than or equal to α (α is a number between 0 and 1, that is to say, to an intensity of 0 to 100%). Note only, that the most sentiment lexicon contain a set of terms with their valences (so if the term is positive, negative or neutral) and the intensity (from 0 to 1). For example the term good can have a positive valence and intensity equal to 1 (that is to say positive at 100%).

3.2 The merge of the two lexicon and replacing of letters

After extracting from the sentiment lexicon the terms with a higher intensity than or equal to α and manually translated these terms in Algerian dialect, we can merge with our dialect lexicon. If the dialect lexicon contains X words and we have extracted Y words from sentiment lexicon. The resultant lexicon will contain X+Y words. However, we observe that this dialect was built for linguistic purposes. Within it, we can find the word under the form of: *dḥâyki* that mean funny, *tbîɛa* which means habit or *ɛawen* means to help. The problem is that no users in social media use the letter "î" / "ɛ" / "ḥ" or the "â" for writing in Algerian dialect. To address this problem, we propose in the first column of Table 2 different replacement which we will proceed. For example, we replace the occurrence of letter *â* by lettre *a*, *ê* by e, š by the two letter « ch », etc. Note only that for proceding to these different replacement letters, we based on two things: 1) the work in Yaghan. (Yaghan 2008), on the transformation of Arabizi letter and 2) our careful analysis of a of a corpus containing a large set of comments from Algerian page on social media.

3.3 Enrichment of lexicon Algerian dialect / French

During our analysis of a user comment corpus of social media, we found that Algerian dialect term could be written in different ways. For example the adjective “blue” can be written in six different ways: *zreq*, *zrek*, *zre9*,*zraq*, *zrak*, *zra9*. The verb “to watch” can be written in eight different ways: *chouf*, *chof*, *c7ouf*, *shouf*, *c7of*, *shof*, *s7ouf*, *s7of*. The name “bird” can be

written in 48 ways, among these we cite :*zawech, zaouech, zewesh, zeouech, zaoac7, zaouesh,…*

The sound		Replacement		Enrichment	
		Letter in the lexicon	Letter used in social medial	Letter in the lexicon	Letter used in social medial
أ	أ	â/ à/ ê	a/e	a	e
				e	a
	أُ	û/ô/ò	ou/	ou	o
	إِ	î	i	i	y
ب		-	-	-	-
ت / ث		t̲	t	-	-
ج		-	-	dj	j
ح		ḥ	h	h	7
خ		x	kh	-	-
د / ذ / ض / ظ		ḍ / d̲	d	-	-
ر		-	-	-	-
ز		-	-	-	-
س / ص		ṣ	s	ss	s
ش		š	ch	ch	sh
ط		ṭ	t	-	-
ع		ε	3	aa	3
				3	aa
غ		ġ	gh	-	-
ف		-	-	-	-
ق		-	-	k	q
				q	k
				k	9
ڤ		-	-	gu	g
ك		-	-	k	q
ل		-	-	-	-
م		-	-	-	-
ن		-	-	-	-
ه		-	-	-	-
و		-	-	wa	oua
				wi	oui
				wi	ui
				oua	wa
				oui	wi
				ui	wi
ي		-	-	-	-

Table 2: **Table representing different letters replacing and letters enrichment of the lexicon**

To manage all the possible entries for the same term, we propose in the second column of Table 2, the different enrichments that we can bring to a letter. For example, we found that users of the Algerian dialect made no difference on social media between the two letters « a » and « e ». The same observation can be done for letters « k » et « q », « wa » et « oua », « ch » et « sh ». We also found a new trend on social media, which is the call to the number replacing the letters. This is done to the letter "h" replaced by the number 7 or the letter "k" and "q" replaced with the number 9. All possible enrichments are represented in the Table. 2.

4 Implementation and results

To implement our approach, we use the programming language JAVA 8 with developing Eclipse environment under the Windows 7 operating system. For the dialect lexicon, we first focus on the nouns, verbs, adjectives and particles. To do this we have extracted the lexicon that we cite before from the net. This lexicon contains 422 verbs, 115 adjectives and 520 names. For the list of particle (pronouns, conjunctions, etc), we can manually generate a list of 86 particles. Regarding the sentiment lexicon, we use SentiWordNet 3.0.0.

This lexicon is largely based on the WordNet ontology. It contains 117 659 rows and 6 columns (POS ID, positive score, negative score, all synonyms of the term and the description of this set). The "POS" column can hold four values: "a" refers to adjectives, "n" to name, "r" to adverb and "v" to the verb. The "ID" column contains the different identifiers of terms in WordNet. The score pos columns and score Neg designate the valence, that is to say the intensity of the term. To have a modular implementation, we cut our approach in three classes: 1) SWN for extracting terms with an intensity of term ≥ α from SentiWordNet and merger with the dialect lexicon. 2) REMP for replacing lexicon letter by letter used in social media. 3) ENRICH for enrichment the obtained lexicon. We present below each of these classes with the intermediate and final results given by the system.

The purpose of the class SWN is to extract terms with an intensity of terms ≥ α from SentiWordNet. To do that, we have to indicate to the application the path of SentiWordNet file and the number α. Note that for the clarity, we have divided SentiWordNet into four categories: adjective, noun, verb and adverb. Within this work, we set α = 1, that is to say that we have extracted the positive and negative terms with 100% intensity. After the application has extracted the most intense terms, we proceed to their translation into French. We find that for verbs, only eight have an intensity equal to100%. We proceed in the same

way for the adjective (where we obtain 27 adjective), names (where we have 33 names) and adverbs (however, in SentiWordNet it has not adverb with an intensity equal to100). At the end of this phase, we collect 429 verbs, 142 adjectives, 553 names and 86 particles. This gives us a total of 1212 peers of terms in Algerian dialect / French. However, these terms are still written with letters such as: â / ô/ ḥ / ε / ... etc. The goal of the next class is to replace these letters with those used in social media.

The purpose of the REMP class is to replace some characters used in the lexicon. We defined in the Table 2, a set of replacement we have found it necessary to our analysis. Nevertheless, the implementation of this class allows changes or extensions of these replacements. We note that at the end of this phase, the number of terms remains unchanged. Nevertheless the letters that have not been used in social media were replaced. However, we have, already seen that the same word can be written in different ways. Now, we present the implementation of the last class, used to enrich the lexicon used.

The Graphical User Inetrface (GUI) that we developp for the class ENRICH is the same that we developp for the class REMP REMP. However instead of replacing letters, the appearance of the letters typed by the user of the application will be enriched by other letters. The list of letters enrichment is in the second column of Table 2. To clarify something, take the case of the adjective *zreq* meaning “blue” that we mentioned in section 3.3. We note that the adjective *zreq* contains two main letters that can be written differently: e and q. In the Table 2 we have indicated that users confused between e and a and between q and k. So we developed an enrichment letter algorithm in combinatorial manner. The different combinations are stored in a dynamic vector of terms. The word insertion process in the vector is remade while it a new terms. Let us clarify that with the case of terme *zrek*: Initially we insert the term *zreq* within the vector. Then the algorithm detects the letters “e” and “q” that must enrich them, respectively by “a”and “k”, so it obtain the new words *zraq ,zrek* and *zrak.* These words, will be insert in the vector. The enrichment is done by respecting the possible combinations, that is to say, only replace the "e" with "a" and only the "q" with "k", and then proceed to two replacements at once. Then, the algorithm detects the "k" letter which can be replaced by the number “9”. it enrich it and inserts the new words *zre9* and *zra9*. This is done to 1212 words in our lexicon.

We illustrate using Table 3, the increasing volume of Algerian dialect lexicon handled during these three phases. The major observation that we can do from this table is that the number of term has largely increased after applying our enrichment algorithme. Note that the total number of words at the beginning was 1144 words. After applaying this three steps, this number increase to 25086 words. However, we observe some noise in final lexicon, eg : when we enrich "3" to "aa", the algorithm proceeds thereafter to the enrichment of each "a" to "e", so we can meet words with a succession of vowels without interest. We can see this, for the term *wa3er* to mean dificult. With enrichment the number 3 by "aa", we can collect some terms with the form of *weaaer* or *waaaar* which mean nothing. So, it would be interested to conduct a filter after the construction of the lexicon, to analyze the collected words and delete words with several repeated letters. Another problem is that we can cited is about the sounds letter that change when it is placed before the "e" or "a". Take for example, the letter "g", pronounced differently before "a" and "e". The confusion and enrichment of "a" to "e" placed neer of these letter could therefore cause errors.

Number of terms	**Verbs**	**Adjectives**	**Nouns**	**Particules**	**ALL**
Lexicon	421	115	520	86	1144
SWN	8	27	33	-	68
Fusion	429	142	553	86	1210
Enrichment	5803	2648	15144	1491	**25086**

Table 3: The number of terms after each step

5 Conclusion and perspectives

As part of our work on the opinion mining and sentiment analysis of users about a given product in social media, we focus on the treatment of arabic

dialects. This treatment has recently attracted a growing interest within the research community. Many works have studied the Arabic dialects but few have targeted the Algerian dialect as use cases. To address this issue we propose in this paper a step by step approach describing a resource construction for the treatment of Arabic dialect, taking as cases of application the Algerian dialect. Our approach follows through three major phases: 1) Extraction and improving a dialect lexicon and extraction of the most intense word from a sentiment lexicon. 2) Merge the two lexicons (dialect end sentiment lexicon) and replace some letters with the most used letters in social media. 3) Enrichment lexicon obtained using spelling variations that appear in social media. We have to keep that we begin with only 1144 words of the Algerian dialect translated into French for reach to 25086 words at the end. However, we intend to improve this lexixon in our future work. This lexicon could be used for many purpose, such as : 1) the identification of the Algerian dialects within comments. 2) Automatic translation from comments written in Algerian dialect into another different language (firstly in French, because it is very close to this dialect). 3) Analysis of feelings of comments written in this dialect. Note also that the algorithme we have implemented to enrich our lexicon, could also be used in other purposes and with other language, such as for the detection of spelling mistakes. However, this work could be improved with the following items:

1. This work is largely based on an Algerian dialect lexicon extracted from the net. It contains a limited number of verbs, adjectives and nouns. However, there are APIs providing for a given term in a given language (possibly the French), the equivalent in Algerian dialect. To enrich this work, it would be interesting to take a dictionary of terms in French, that is to say containing all the words of French. Our program would then call the API for each term. This will allow us to start with a larger lexicon.

2. Automating the analysis part, so automatically produce table 2. It would be interesting to have a module that analyzes the most used letters in the comments written in dialect on social media. This module also analyze the letter would extract confused. That is to say, we can say that social media users confused between "a" and "e" in their messages. This statement may be doing automatically with results.

Treatment lexicon obtained after enrichment in order to increase accuracy and eliminate noise. That is to say, delete the words that are left with several repeating vowels or avoid using "a" and "e" in the same way with the letters “g” or “c”.

Acknowledgments

We would like to thank different researcher for having send us their lexicons and corpora. Between them: Rayan Cottrel for its code switched corpus, Gilbert Badaro for ArSenl lexicon, Pierre charron for NRC corpus, Eshrag Refaee for its Twitter annotated corpus and Khaled Elmiman for its Multi dialect text corpora.

References

Abdul-Mageed M, Diab M, Kübler S (2014) SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language 28 (1):20-37

Al-Sabbagh R, Girju R A supervised POS tagger for written Arabic social networking corpora. In: KONVENS, 2012a. pp 39-52

Al-Sabbagh R, Girju R YADAC: Yet another Dialectal Arabic Corpus. In: LREC, 2012b. pp 2882-2889

Almeman K, Lee M Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In: Communications, signal processing, and their applications (iccspa), 2013 1st international conference on, 2013. IEEE, pp 1-6

Bouamor H, Habash N, Oflazer K A Multidialectal Parallel Corpus of Arabic. In: LREC, 2014. pp 1240-1245

Boujelbane R, BenAyed S, Belguith LH (2013) Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model. ACL 2013:88

Cotterell R, Renduchintala A, Saphra N, Callison-Burch C An algerian arabic-french code-switched corpus. In: Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, 2014. p 34

Diab MT, Al-Badrashiny M, Aminian M, Attia M, Elfardy H, Habash N, Hawwari A, Salloum W, Dasigi P, Eskander R Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In: LREC, 2014. pp 3782-3789

Duh K, Kirchhoff K Lexicon acquisition for dialectal Arabic using transductive learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing, 2006. Association for Computational Linguistics, pp 399-407

Graff D, Buckwalter T, Jin H, Maamouri M Lexicon Development for Varieties of Spoken Colloquial

Arabic. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006. Citeseer, pp 999-1004

Guellil I, Boukhalfa K Social big data mining: A survey focused on opinion mining and sentiments analysis. In: Programming and Systems (ISPS), 2015 12th International Symposium on, 2015. IEEE, pp 1-10

GUELLIL, Imane, AZOUAOU, Faiçal, ABBAS, Mourad, *et al.* Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic. In : *Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation (co-located with EAMT2017)*. 2017b.

GUELLIL, Imène et AZOUAOU, Faiçal. Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect. In : Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on. IEEE, 2016. p. 724-731.

GUELLIL, Imène et AZOUAOU, Faiçal. ASDA: Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique. *arXiv preprint arXiv:1707.08998*, 2017a.

Habash N, Rambow O Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. In: International Symposium on Computer and Arabic Language (ISCAL), Riyadh, Saudi Arabia, 2007.

Hamdi A, Nasr A, Habash N, Gala N POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools. In: ANLP Workshop 2015, 2015. p 59

Harrat S, Meftouh K, Abbas M, Smaili K (2014) Building Resources for Algerian Arabic Dialects. Corpus (sentences) 4000 (6415):2415

Jehl L, Hieber F, Riezler S Twitter translation using translation-based cross-lingual retrieval. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, 2012. Association for Computational Linguistics, pp 410-421

Maamouri M, Bies A, Buckwalter T, Diab M, Habash N, Rambow O, Tabessi D Developing and using a pilot dialectal Arabic treebank. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, 2006.

Meftouh K, Bouchemal N, Smaïli K A study of a non-resourced language: an Algerian dialect. In: SLTU, 2012. pp 125-132

Sadat F, Kazemi F, Farzindar A Automatic identification of arabic dialects in social media. In: Proceedings of the first international workshop on Social media retrieval and analysis, 2014a. ACM, pp 35-40

Sadat F, Mallek F, Sellami R, Boudabous MM, Farzindar A Collaboratively Constructed Linguistic Resources for Language Vari-ants and their Exploitation in NLP Applications–the case of Tunisian Arabic and the Social Media. In: Workshop on Lexical and Grammatical Resources for Language Processing, 2014b. p 102

Shoufan A, Al-Ameri S Natural Language Processing for Dialectical Arabic: A Survey. In: ANLP Workshop 2015, 2015. p 36

Yaghan MA (2008) "Arabizi": A Contemporary Style of Arabic Slang. Design Issues 24 (2):39-52

Zaghouani W Critical survey of the freely available Arabic corpora. In: Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC, 2014. pp 1-8

Zaidan OF, Callison-Burch C (2014) Arabic dialect identification. Computational Linguistics 40 (1):171-202