Bilingual Lexicon for Algerian Arabic Dialect Treatment in Social Media

Imane GUELLIL
Ecole Supérieure d'Informatique ESI Alger
Ecole préparatoire des sciences et techniques (EPSTA) Alger
i_guellil@esi.dz
i.guellil@epsta.dz

Faical AZOUAOU
Ecole Supérieure d'Informatique ESI Alger
f_azouaou@esi.dz

Abstract

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. The works on Arabic dialects can be classified into four categories: basic analysis, resource construction, identification of Arabic dialects and semantic analysis on Arabic dialects. In this paper, we address the problem of Arabic lexicon construction. More specifically, we focus on Algerian dialect which is spoken by more than 40 million people. Our approach consists of three distinct steps: 1) Constructing a dialect lexicon and extracting the most intense words from a sentiment lexicon; 2) Merging two lexicons (a dialect and a sentiment lexicon) and replacing some letters with the most used letters in social media; 3) Enriching a lexicon that takes into account spelling variations of words specific to a social media domain. As a result, we ended up expanding the original lexicon of 1144 words to a new lexicon 25086 words.

1 Introduction

Arabic dialects are grouped into six categories: EGYPTIAN (Egypt), LEVANTINE (Syria and Palestine), GULF (the gulf countries), IRAQI (Iraq) and MEGHREBI (North African countries) and others (Zaidan and Callison-Burch 2014). The earlier work on these dialects focused on four problems out lighted below: basic analysis of dialects, resource construction, identification of Arabic dialects, and, finally, semantic analysis on Arabic dialects (Shoufan and Al-Ameri 2015). Several works on resource construction(both lexicon and corpus) focused on the Iraqui dialect (Graff et al. 2006), and Tunisian dialect (Boujelbane et al. 2013.

However, to the best of your knowledge, only limited work focused on the Algerian dialect. Moreover, the Algerian dialect is considered to be an under-resourced language (Meftouh et al 2012). To bridge this gap, we develop a bilingual lexicon between the Algerian dialect and other languages. This lexicon will be used for a dialect identification task and for the automatic translation of this dialect to another language.

2 Related work

In this section, we present the related work on lexicon construction. For the Egyptian dialect, Diab et al. (2014) focused on the construction of **Tharwa**. This lexicon contains 73 348 words, and is between the MSA and the Egyptian dialect.

For the Levantine (LEV) dialect, Duh et al. (2006) constructed the LEV/MSA lexicon. However, the authors found that the LEV differs significantly from the MSA.
For the Tunisian dialect (TUN), Boujelbane et al. (2013) focused on creating resources (both lexicon and corpus) for the translation of the Tunisian dialect. Hamdi et al. (2015) developed the POS tagger for a given language using the resources of another.
For IRAQI, Graff et al. in (2006), focused on six different lexicons of the Iraqi dialect. The result of this work contains a complete set of pronunciations, morphology, the POS tags and annotations in English for 120 000 words. For the Algerian dialect Harrat et al. (2014) developed a translation between the Algerian dialect and classical Arabic (MSA). The work focused on two types of the Algerian dialects: (ALGR) which is the dialect of Algiers, and (ANB), which is the dialect of Annaba (a city in Algeria). The authors built two lexicons: the first is between the MSA and ALGR (10 790 words), and the second is between the MSA and ANB (9688 terms).

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

The main contribution of this work is the construction and enrichment of a bilingual lexicon between the Algerian dialect, mainly Algiers, and French. This work represents one component that is focused on opinion mining and sentiment analysis (Guellil et al. 2015). In our lexicon we integrate sentiment bearing terms. To construct the lexicon, we follow three steps: 1) We extract and improve a dialect lexicon by integrating sentiment bearing terms into the lexicon. 2) We merge a dialect and a sentiment lexicon, and replace some characters with the most used characters in social media. 3) We enrich the resulting lexicon with spelling variations that appear in social media. Figure.1 shows the transition between these three steps.

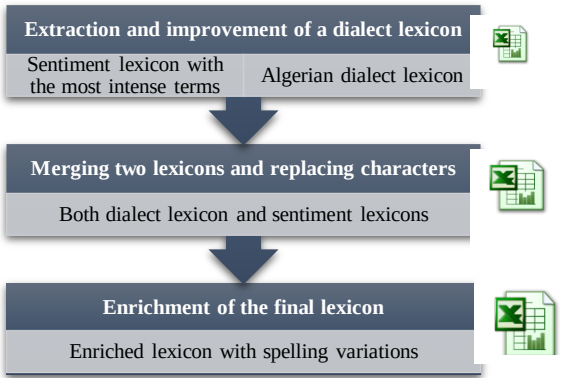

Figure 1: The process of construction and enrichment of a bilingual lexicon between the Algerian dialect and French.

3.1 Extraction and Manual Improvement of a Dialect Lexicon

We first extract a novel lexicon (to the best of our knowledge) between the Algerian dialect and French. It contains the translation between the most used verbs, nouns and adjectives in the Algerian dialect. Then, we add particle parts to this lexicon: *fi* (in),*Fouk* (on) or *nchalah* (if god want), that are used a lot by Algerians. Let us simply note that the generation of particles is inspired by earlier work of Hamdi et al. (2015) for the Tunisian dialect (TUN). Finally, we incorporate sentiment bearing terms related to feelings or sentiments such as love, hate, sadness, etc.

3.2 Dialect and Sentiment Lexicon Merge

After extracting sentiment bearing terms with higher intensity or equal that so α threshold, we manually translate these terms to the Algerian dialect. We then merge the translated sentiment lexicon with the dialect lexicon. If the dialect lexicon contains X terms and we have extracted Y terms from the sentiment lexicon. The resultant lexicon will contain X+Y words.

3.3 Final Enrichment of the Algerian dialect – French Lexicon

During our analysis of a comment corpus on social media, we found that the Algerian dialect terms could be written in different ways. For example, the adjective “blue” can be written in six different ways: *zreq, zrek, zre9,….* The verb “to watch” can be written in eight different ways: *chouf,c7ouf.*
To manage all the possible variations for the same term, we propose spelling enrichments as shown in the second column of Table 1. For example, we

found that speakers of the Algerian dialect made no difference on social media between the letters « a » and « e ». The same observation can be done for the letters « k » and « q », « wa » and « oua », « ch » et « sh ». Our final spelling enrichments are presented in Table. 1.

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		-	-	-	-
ث / ت		ṯ	t	-	-
ج		-	-	dj	j
ح		ḥ	h	h	7
خ		x	kh	-	-
ذ / د ظ / ض		ḍ / ḏ	d	-	-
ر		-	-	-	-
ز		-	-	-	-
ص / س		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 1: Spelling enrichments in the final lexicon.

4 Implementation and Results

Table 2 shows the increasing volume of the proposed Algerian dialect lexicon at three stages of its construction. We demonstrate how the size of the lexicon is largely increasing after applying several enrichment steps.

Number of terms	Verbs	Adjectives	Nouns	Particules	ALL
Lexicon	421	115	520	86	1144
SWN	8	27	33	-	68
Fusion	429	142	553	86	1212
Enrichment	5803	2648	15144	1491	**25086**

Table 2: The size of the dialect lexicon at different stages of its construction.

5 Conclusions and Future Work

In this paper, we propose an approach for constructing a new resource for the Algerian dialiect. Our approach involves three major steps: 1) extracting and improving the initial Algerian dialect lexicon; 2) Merging the dialect lexicon with the translated sentiment lexicon ; 3) Enrching the final lexicon using spelling variations specific tosocial media. We started with 1144 terms in the original Algerian dealect lexicon, and we ended up with 25086 terms in the final lexicon. In future, we plan to 1) automate the analysis part to automatically produce table 1, 3) verify the final lexicon to increase accuracy and eliminate noisy entries.

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 ap-

plications (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

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

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.

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