Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, Kamel Smaili

Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus

Karima Meftouh Badji Mokhtar University Annaba, Algeria Karima.meftouh@univ-annaba.org

Salima Harrat ENSB [*], ESI [†] Algiers, Algeria slmhrrt@gmail.com

Salma Jamoussi MIRACL[‡] Pole Technologique de Sfax, Tunisia jamoussi@gmail.com

Mourad Abbas CRSTDLA[§] Algiers, Algeria m_abbas04@yahoo.fr

Kamel Smaili Campus Scientifique LORIA Nancy, France smaili@loria.fr

Abstract

We present in this paper PADIC, a Parallel Arabic DIalect Corpus we built from scratch, then we conducted experiments on cross-dialect Arabic machine translation. PADIC is composed of dialects from both the Maghreb and the Middle-East. Each dialect has been aligned with Modern Standard Arabic (MSA). Three dialects from Maghreb are concerned by this study: two from Algeria, one from Tunisia, and two dialects from the Middle-East (Syria and Palestine). PADIC has been built from scratch because the lack of dialect resources. In fact, Arabic dialects in Arab world in general are used in daily life conversations but they are not written. At the best of our knowledge, PADIC, up to now, is the largest corpus in the community working on dialects and especially those concerning Maghreb. PADIC is composed of 6400 sentences for each of the 5 concerned dialects and MSA. We conducted cross-lingual machine translation experiments between all the language pairs. For translating to MSA we interpolated the corresponding Language Model (LM) with a large Arabic corpus based LM. We also studied the impact of language model smoothing techniques on the results of machine translation because this corpus, even it is the largest one, it still very small in comparison to those used for translation of natural languages.

[*] École Normale Supérieure Bouzareah.

[†] École Supérieure d'Informatique.

[‡] Multimedia, InfoRmation systems and Advanced Computing Laboratory.

[§] Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe.

1 Introduction

Recently, in addition of translating MSA (Modern Standard Arabic), a new challenging issue emerges: How to deal with the translation of Arabic dialects? Few years ago, some works have been proposed to process Arabic dialects and more specifically those of Middle-East. These works concerned building lexicon, analyzing text morphology, POS tagging, diacritization, etc, (Kilany et al., 2002; Kirchhoff et al., 2003; Habash and Rambow, 2006; Chiang et al., 2006; Habash et al., 2013; Elfardy and Diab, 2013; Pasha et al., 2014; Harrat et al., 2014). In the context of Machine Translation some Arabic dialects have started receiving increasing attention (Sawaf, 2010; Zbib et al., 2012; Salloum and Habash, 2013).

Number of researchers have exploited the NLP tools dedicated to MSA to develop their Machine Translation (MT) systems for Arabic dialects, considered as under-resourced languages. Ridouane and Karim (2014) used tools designed for MSA and adapted them to Moroccan dialect in order to build a translation system from MSA to Moroccan, by combining a rule-based approach and a statistical approach. Sawaf (2010) built a hybrid AD-English MT system that uses a MSA pivot approach. In this approach, AD is transferred into MSA using character-based AD normalization rules. In addition an AD normalization decoder that relies on language models, an AD morphological analyzer, and a lexicon were employed to achieve the translation task. Similarly, Salloum and Habash (2012) presented Elissa , an MT system from AD to MSA which employed a rule-based approach that relies on morphological analy-

sis, morphological transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences. Elissa handles Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic. Zbib et al. (2012) used crowdsourcing to build Levantine-English and Egyptian-English parallel corpora. They selected dialectal sentences from a large corpus of Arabic web text, and translated them using Amazon's Mechanical Turk platform. They used this data to build dialectal Arabic MT systems, and find that small amounts of dialectal data have a dramatic impact on translation quality.

Multidialectal Arabic parallel corpora do not exist, the first and the unique such corpus was presented in (Bouamor et al., 2014). It is a collection of 2000 sentences in MSA, Egyptian, Tunisian, Jordanian, Palestinian and Syrian, in addition to English. The sentences were selected from the Egyptian part of the Egyptian-English corpus built by Zbib et al. (2012).

In this paper, we present PADIC, a corpus composed of 5 Arabic dialects, each of them contains 6400 sentences. Each dialect is aligned at the sentence level with the four other dialects and also with MSA. In this paper, we highlight machine translation results obtained for all the pairs of dialects contained into PADIC. The reminder of this paper is organized as follows, in section 2 we give some differences that distinguish MSA from its dialects. Section 3 describes the processes we used to build PADIC. Finally, we present in section 4 experiments of machine translation between several pairs of dialects and MSA. We also show the impact of the smoothing techniques over the BLEU scores due to the size of the training corpora.

2 Main differences between modern standard Arabic and its dialects

MSA is characterized by a complex morphology and a rich vocabulary. It is an inflexional and agglutinative language. We recall that, compared to English, an Arabic word (or more rigorously a lexical entry) can sometimes correspond to a whole English sentence.

The differences that distinguish dialects from MSA are at the morphological, lexical and syntactical levels. Because it is difficult for a non-Arabic to understand these differences, let us give some examples. In Maghreb, the phrase ما نكتبش "*mā nktbš*" (*I dont write*) is the negation of the word نكتب (*I write*). While in MSA, the negation form is expressed by one of the two function words لا "*lā*" or ما "*mā*" . Consequently, *I don't write* ما نكتبش is written as follows in MSA: لا أكتب "*lā ʾktb*" . For the Maghrebian dialect, such as in the previous example, the morpheme ش /*š*/ is added to the end of the stem نكتب, and the negation marker "ما *mā*" is inserted in the beginning of the phrase[1]. Between MSA and the above dialect, morphologically, we have the same stem كتب "*ktb*" , however for the Maghrebian dialects, the affixes have been changed and a new one has been added. Lexically, most dialectal words are taken from MSA, however many foreign words (verbs and nouns) have been introduced in the daily spoken communications. Maghrebian dialect speakers often use foreign verbs with some modifications; the expressions: شرجاها "*šarğāhā*" and يشرجيها "*yšarğīhā*" for respectively *he charged it* and *he charges it* are noticed as two single words but in reality are two sentences, formed by concatenating the morphemes ها /*hā*/ (the object) and ي /*y*/ (the subject) to the verb شرجا "*šarğā*" *to charge* .

Syntactically, the Verb-Subject-Object order (VSO) is common in MSA and so is (SVO), but (OVS) and (OSV) are also correct even they are not widely used. Nevertheless, in some dialects (SVO) is more used than (VSO) such as in Levantine Arabic. In other dialects as Maghrebian, (VSO) is more preferred. Up to now, no one is able to give an answer to: *which is the closest dialect to MSA?*. It is then necessary to go through a comparative study of these dialects to objectively answer this question. In fact, some old expressions of classical Arabic are still used by Maghrebian people and no longer used in the Arabian Peninsula. Inversely, other aspects of MSA are preserved in the Arabian Peninsula, such as Tanween (to mark indefiniteness) but not used in North Africa at all.

[1]The morpheme ش "*š*" is the abridged dialectal form of the MSA word شيء "*šayʾ*" "thing". Ex: the origin of the word ما نكتبش "*mā nktbš*" is ما نكتب شيء for لا أكتب شيئا "I dont write anything".

Necessity of adopting writing rules

MSA is a natural language with linguistic rules and a typographic system of writing, dialects do not have any standardized set of rules. In fact, there is no reason to write dialects which are usually spoken in daily conversations. But a new phenomenon arises with social networks: people are free to write whatever they want and express their opinions such as they speak, it means in their dialect. Accordingly, they write dialectal words just as they are pronounced. For instance, to write *tell him*, one would write it, just as he heard it: "*qūllu*" قولّو while the right expression is "*qūl lū*" قول لو (original expression in MSA is قل له).

This freedom of writing pushed us to adopt some writing rules for standardizing our corpus. In our work each dialectal word is written by adopting MSA rules, that means if a dialectal word does exist in MSA, it must be written such as in MSA, otherwise the word is written as it is uttered. In this last case, we could extract the phonetic directly form from its orthographic representation, which will be necessary in the frame of the ultimate goal of this study which is Speech-To-Speech translation.

In Arabic dialects, some foreign words contain non Arabic phonemes, such as /g/ which could be written either with /غ/ or /ج/ such as for the English words *English* and *Ghana* which are respectively written انجليزي and غانا.

However, the few dialectal texts retrieved from the web constitute a big challenge to researchers. This is not only because of the non standardization of orthographies, but also due to the absence of diacritics as it is the case for MSA and all its dialects (Harrat et al., 2013). In social networks, Arabic dialects are written in different ways, sometimes in Arabic script, sometimes in Latin one and in some cases such as a mixture of letters and some specific numbers. For example, the number 3 is used to represent the phoneme /ع/ because of the similarity between 3 and /ع/. In Table 1, we address some Arabic letters and the adopted Arabic numbers used to represent them. Note the similarity in the form between the letters and the numbers.

To illustrate the different ways of writing dialects in social networks, in Table 2, some examples of Algerian sentences are given.

Table 1: Numbers adopted to represent some Arabic letters when Latin grapheme are used to write Arabic

Example	Arabic number	Arabic letter
tbarra3	3	/ع/
fra7	7	/ح/
sou9	9	/ق/

3 Building a parallel corpus

It is well known that parallel corpora are the foundation stone of several natural language processing tasks, particularly cross-language applications such as machine translation, bilingual lexicon extraction and multilingual information retrieval. Building this kind of resources is a challenging task especially when it deals with under-resourced languages (Skadiņa et al., 2010). The problem is much deeper with the Arabic dialects which are used by a huge number of people only in daily oral communication. Moreover, they are not taught in schools and are absent from formal written communications. This makes building dialectal resources automatically almost impossible. The few available texts in social networks, usually produced by less educated Arab people are not homogeneous and suffer from the varieties in orthography, due to the absence of writing rules. In addition, some Arabic dialects are written by using Latin graphemes by a slice of Arabic societies. The reason is that Arabic language was not widely supported by devices. Consequently, Arab people have been used to this situation by using Latin graphemes.

For all the aforementioned reasons, crawling the web is not a solution to build a parallel corpus, thus, we decided to build it from scratch.

PADIC: A New Parallel Arabic Dialect Corpus

The approach we used to build PADIC is almost similar to that of Bouamor et al. (2014) except that in our case, we started from scratch and we are still working on the development of all the necessary tools. We should note that the particularity of our corpus is that it is composed also of Maghrebian dialects that present difficulties to collect and process since they are mixture of several languages (Arabic,

Table 2: Different ways of writing dialects on Facebook.

Cases of writing dialects	Dialectal sentences	Equivalent in English
Written with Arabic letters	واش راك خويا	How do you do my brother?
Written only with Latin letters	Wachrak khouya?	How do you do my brother?
Written with a mixture of Latin letters and Arabic numbers	rabi ya7fedek	God protect you
A sentence that contain both Arabic and French words	*Et bien* hakda rak f9edt'houm *les deux*	So you have lost both of them

French, Berber, ...). Also, because they are not much used on the Web and when it is the case, people use generally Latin script for writing, as we mentioned it in section 2.

A preliminary study of the PADIC corpus was given in (Harrat et al., 2015). The goal of this work is to focus more on Statistical Machine Translation experiments from MSA to dialectal Arabic and vice versa. This work in turn is part of a big and challenging project, a Speech to Speech Translation system that we are working on. Challenging not only because speech recognition and speech synthesis are involved, but also because of the lack of dialectal Arabic parallel corpora.

Five dialects, in addition to MSA, are concerned by this study: Annaba's dialect (ANB), spoken in the east of Algeria, Algiers's dialect (ALG), used in the capital of Algeria, Sfax's dialect (TUN) spoken in the south of Tunisia, Syrian and Palestinian dialects (SYR) and (PAL) which are spoken in Damascus and Gaza respectively. [2]

ANB corpus was created by recording different conversations from every day life, whereas, for ALG, we used the recordings corresponding to movies and TV shows which are often expressed in the dialect of Algiers. Then we transcribed both of them by hand. To increase the size of the two corpora, we translated each of them into the other. Afterwards, these two corpora have been translated into MSA.

MSA was then used as a pivot language to get other dialectal corpora. To do that, we translated the MSA corpus to TUN, SYR and PAL. The Tunisian corpus was produced by 20 native speakers. Each one was responsible of translating almost 320 sentences from MSA to TUN. Speakers have very slight differences in their spoken languages. All of them are from the south of Tunisia where people tend to use Arabic words rather than French words as it is the case in the north of the country. In fact, the dialect used in the south is closer to MSA than that used in the north of Tunisia. Syrian and Palestinian corpora were created in the same way by respectively two native speakers. Finally, we get a multi-dialectal parallel corpus PADIC composed of the five aforementioned dialects, in addition to MSA. PADIC is made of 6400 parallel sentences, for which we present some statistics in Table 3.

Table 3: PADIC statistics

Corpus	#Distinct words	#Words
ALG	8966	38707
ANB	9060	38428
TUN	10215	36648
SYR	9825	37259
PAL	9196	39286
MSA	9131	40906

The MSA part contains 40906 words including 9131 different words. PADIC includes an average of 37500 words for one a dialect with a vocabulary which does not exceed 10250 words. The average number of words in a dialectal sentence is of 6 while it is of 7 for MSA. The shortest sentence in the corpus is composed of 4 words and the longest one contains 25 words.

We give in Table 4 examples of parallel sentences from PADIC. Even if we do not read Arabic at all, we can notice that some words have the same form in several dialects, while others are completely dif-

[2] The only argument in the choice of these dialects and not others is that the authors of this paper are from Algeria and Tunisia and for the two others we asked kindly colleagues from Syria and Gaza to help us to translate a MSA corpus into these two dialects without any financial compensation. Translating the corpus into Moroccan dialect is under work.

Table 4: Examples of parallel sentences from PADIC

Lang.	Sentence
ALG	جوزت ايامات روعة ما ننساهاش طول حياتي
ANB	عقبت ايامات على الكيف ما ننساهاش طول عمري
TUN	عديت ايامات حلوة ما ننساهاش طول عمري
SYR	مضيت أيام حلوة عمري ما بنساها
PAL	قضيت أيام جميلة مش حانساها طول عمري
MSA	أمضيت أيام جميلة لن أنساها طول عمري
EN	I spent beautiful days I will never forget throughout my life.
ALG	خدمت في وحد السبيطار قريب من دارنا الحمد لله راني لاباس و عايش مع بابا
ANB	خدمت في وحد السبيطار ڤريب من دارنا الحمد لله أموري مليحة و عايش مع بابا
TUN	خدمت في سبيطار قريب من الدار الحمد لله أموري باهية و عايش مع بابا
SYR	إشتغلت بمشفى قريب من بيتي الحمد لله أموري ميسرة و عايش مع أهلي
PAL	إشتغلت في واحد من المستشفيات القريبة من بيتي الحمد لله أموري متيسرة و عايش مع أبوي
MSA	عملت في أحد المستشفيات القريبة من بيتي الحمد لله أموري متيسرة و أعيش مع والداي
EN	I worked in a hospital near my home. Praise be to God, everything is fine and I live with my parents

ferent.

4 Experiments on Machine Translation of Arabic dialects

In the following, we present several experiments in machine translation in both sides between all the combinations of dialect pairs. We conduct also experiments of machine translation between these dialects and MSA also in both sides.

All the MT systems we used are phrase-based (Koehn et al., 2007) with default settings: bidirectional phrase and lexical translation probabilities, distortion model, a word and a phrase penalty and a trigram language model. We have not used a larger language model because PADIC is not suitable for large ngrams. We used GIZA++ (Och and Ney, 2003) for alignment and SRILM toolkit (Stolcke, 2002) to compute trigram language models. Since the parallel corpus is small, we experimented the Kneser-Ney and Witten-Bell smoothing techniques hoping to identify the one which best fits. The results conducted on a test set of 500 sentences are presented in terms of BLEU, TER and METEOR in Tables 5, 6 and 7 respectively.

Because it is difficult to increase the size of PADIC, we decided to interpolate the corresponding language models by larger Arabic corpora when the target language is MSA. For this purpose, we used two MSA corpora: Tashkeela[3] and LDC Arabic Treebank (Part3,V1.0) (Maamouri et al., 2004). Unfortunately, the results of the interpolation did not outperform those of Table 5.

5 Discussion

5.1 Impact of the smoothing techniques on BLEU

Several conclusions can be presented regarding results of the Table 5. First of all, for a small parallel corpus, it seems that the smoothing technique has an impact on the BLEU scores. A difference of almost 2 points has been observed for translating from ANB to ALG. But, we can not generalize by affirming that one smoothing technique is definitely better than another. We have not calculated the significance of this difference because our corpus is too small, consequently we can not have several small test corpora in order to perform significance tests. In order to have an idea about the impact of the smoothing technique on the results and to have a scale comparison of the BLEU for small corpora we did some

[3] A collection of classical Arabic books from an online library available on http://sourceforge.net/project/tashkeela

Table 5: BLEU score of Machine Translation on different pairs of languages using two smoothing techniques

	Target											
	ALG		**ANB**		**TUN**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	61.06	60.81	9.67	9.36	7.29	7.95	10.61	10.14	15.1	14.64
ANB	67.31	65.55	-	-	9.08	8.64	7.52	7.95	10.12	9.84	14.44	13.95
TUN	9.89	9.48	9.34	9.01	-	-	13.05	12.93	22.55	22.21	25.99	25.21
SYR	7.57	7.50	7.50	7.64	13.67	13.23	-	-	26.60	25.74	24.14	22.96
PAL	11.28	10.67	9.53	9.15	17.93	16.64	23.29	23.07	-	-	40.48	39.76
MSA	13.55	13.05	12.54	11.72	20.03	20.44	21.38	20.32	42.46	41.37	-	-

Table 6: TER score (in %) of Machine Translation on different pairs of languages using two smoothing techniques

	Target											
	ALG		**ANB**		**TUN**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	21.41	21.75	75.17	76.37	79.54	79.51	69.63	70.75	65.63	66.85
ANB	17.12	17.81	-	-	74.83	75.62	79.13	79.13	69.10	70.26	67.40	68.47
TUN	71.10	71.76	73.13	73.71	-	-	66.03	66.55	51.20	51.57	50.85	51.30
SYR	76.89	77.67	76.89	77.67	66.54	67.91	-	-	32.28	33.24	52.81	53.59
PAL	71.43	72.51	72.25	73.47	58.51	59.65	32.44	33.86	-	-	36.74	36.87
MSA	67.02	67.91	68.94	70.16	57.18	57.28	56.60	57.08	34.00	34.66	-	-

Table 7: METEOR score of Machine Translation on different pairs of languages using two smoothing techniques

	Target											
	ALG		**ANB**		**TUN**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	0.452	0.450	0.181	0.178	0.161	0.164	0.202	0.199	0.222	0.218
ANB	0.472	0.464	-	-	0.172	0.172	0.156	0.159	0.196	0.194	0.200	0.200
TUN	0.186	0.183	0.182	0.182	-	-	0.203	0.203	0.261	0.260	0.268	0.266
SYR	0.155	0.154	0.159	0.157	0.195	0.190	-	-	0.359	0.356	0.259	0.256
PAL	0.189	0.185	0.187	0.183	0.229	0.225	0.365	0.360	-	-	0.341	0.339
MSA	0.205	0.203	0.201	0.199	0.242	0.245	0.247	0.247	0.359	0.356	-	-

experiments on a small parallel Arabic-English corpus extracted from WMT. We took small corpora in order to be approximatively in the same context such as with PADIC. We used several small training parallel corpora of $20K$, $50K$, $120K$ and $150K$ parallel sentences which will be denoted respectively S_2, S_5, S_{12} and S_{15}. For each training corpus we performed 20 experiments on 20 different small test corpora (500 sentences such as for the dialects). The results are presented in Table 8.

In Table 8, *Min*, *Max*, *E[X]*, σ^2 represent respectively the *minimum*, *maximum*, *mean* and *variance* of the corresponding distribution of BLEU according to the used smoothing technique. While σ_{XY} and *p-value* correspond respectively to the *covariance* and the *p-value* of the two distributions. The statistical test used is T-test after checking that the two distributions follow a Gaussian law. The hypothesis H_0 is that the two distributions are similar (in terms of mean).

According to these different results, it seems that the results obtained by the first or the second smoothing techniques are not distinguishable since for each training corpus and for 20 different tests, the results are equivalent in terms of *minimum*, *maximum*, *mean* and *variance* BLEU values. Furthermore, the covariance is positive for all the experiments which would mean that the two distributions are linearly dependent. The *p-value* whatever the training corpus is greater than the α risk set to 0.05 which means that there is at least a risk of 33% to accept the alternative hypothesis H_1. In conclusion, unfortunately even if there is a difference between the results of BLEU according to the used smoothing techniques, the difference is not significant.

5.2 Cross-translation results comparison

High score of translation has been achieved between ANB and ALG in both sides. This result is natural since these two dialects are spoken in the same country and share up to 60% of words. Almost the same observation is made for the pair SYR and PAL since these two dialects belong to the same language family (Levantine).

Another interesting and expected result is BLEU score between MSA and dialects. In fact, the highest one is related to PAL (for both sides) showing that this dialect is the closest to MSA. Most surprising results are those relative to SYR and TUN. It seems that it is easier to translate TUN to MSA than SYR to MSA. Also, translating from MSA to TUN gives better results than from MSA to the Algerian dialects. In the symmetric side of translation we get the same scale of results. This definitely shows the closeness of TUN dialect to MSA in comparison to the Algerian dialects. The same conclusions can be inferred from the results in terms of TER or METEOR. Also, it seems that the smoothing technique has no impact on both scores. The differences are almost negligible.

We can notice that the values of BLEU are weak in comparison to what the community get usually for large training corpora. This is obviously due to the weak size of the training corpora. But when we compare the scale values of BLEU for dialects to those achieved for English-Arabic (Table 8) which have been performed with small training corpus, we notice that those obtained for dialects are higher. This is probably due to the fact that, even if dialects are very different, nevertheless there is a strong relationship between them. For instance the experiment performed on the corpus S_2, the smallest value of BLEU is 4.25 and the highest is 9.56 while for the worst results of translation (from Syrian to Algerian) the minimum value is 7.57. Knowing that for this comparison, the training corpus S_2 (English-Arabic) is 3 times larger than the one used for the dialect.

6 Conclusion

In this paper, we first presented PADIC a parallel corpus containing five dialects from Maghreb and middle-east. PADIC has been built from scratch because there is no standard resources due to the kind of theses languages which are only used for conversations and not for writing. Then, we presented experiments on cross-dialectal Arabic machine translation. In the best of our knowledge, this is the first work on machine translation of dialects from both Maghreb and Middle-East and also the largest existing parallel Arabic dialect corpora. On the limited corpora of 6400 parallel sentences we built, we achieved some interesting results.

Due to the small size of the corpus, we analyzed the impact of the language model on the process

Table 8: Statistics on machine translation with small training corpus and by varying the smoothing techniques of the language model.

	KN				WB					
Corpus	Min	Max	$E[X]$	σ^2	Min	Max	$E[X]$	σ^2	σ_{XY}	*p-value*
S_2	4.25	9.56	6.64	2.47	4.1	8.97	6.43	2.23	2.33	0.33
S_5	5.15	11.75	8.15	3.26	5.18	11.32	7.99	2.92	3.16	0.35
S_{12}	5.94	14.38	9.58	4.62	5.95	14.13	9.35	4.32	4.45	0.36
S_{15}	6.13	14.39	9.91	4.85	6.19	14.27	9.74	4.72	4.75	0.39

of machine translation by varying the smoothing techniques and by interpolating it with a larger one trained on well known corpora. Unfortunately the results are not significant even if in some cases we get some improvements.

The best results of translation are achieved between the dialects of Algeria. This is not a surprising result since they share a large part of the vocabulary. And even if the size of the training corpus is weak, we noticed that the BLEU is very high. The performance of machine translation between Palestinian and Syrian are relatively high in accordance to the size of the corpus. This could be explained by the closeness of the two regions. The worst result is achieved between Syrian and Algerian dialects which are, in fact, very different since the Algerian borrowed a lot of French words which do not exist obviously in the Syrian dialect. Concerning MSA, the best results of machine translation have been achieved with Palestinian dialect. This means that the two languages are very close since they share a large number of words.

Our future work consists in extending PADIC to other dialects such as Moroccan in order to have the dialects of the three countries of Maghreb and then we will work on using the large existing corpora of MSA to rewrite part of them into dialects.

Acknowledgement This work has been supported by PNR (Projet National de Recherche of Algerian research Ministry).

References

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference, LREC-2014*, pages 1240–1245.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Heba Elfardy and Mona Diab. 2013. Sentence Level Dialect Identification in Arabic. In *ACL (2)*, pages 456–461.

Nizar Habash and Owen Rambow. 2006. Magead: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal arabic. In *Proceedings of NAACL-HLT*, pages 426–432, Atlanta, Georgia.

Salima Harrat, Mourad Abbas, Karima Meftouh, and Kamel Smaili. 2013. Diacritics restoration for arabic dialect texts. In *Proceedings of 14th Annual Conference of the International Communication Association (Interspeech)*, pages 125–132.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaili. 2014. Building resources for algerian arabic dialects. In *Proceedings of 15th Annual Conference of the International Communication Association (Interspeech)*, pages 2123–2127.

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, and Kamel Smaili. 2015. Cross-dialectal arabic processing. In *Computational Linguistics and Intelligent Text Processing, 16th International Conference, CICLing 2015 proceeding, part 1*, pages 620–632, April.

Hanaa Kilany, H. Gadalla, Howaida Arram, A. Yacoub, Alaa El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. In *LDC catalog number LDC99L22*.

Katrin Kirchhoff, Jeff Bilmes, Sourin Das, Nicolae Duta, Melissa Egan, Gang Ji, Feng He, John Hopkins, Daben Liu, Mohamed Noamany, Pat Schone, Richard Schwartz, and Dimitra Vergyri. 2003. Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop. In *Proc.*

IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, pages 344–347.

Philipp Koehn, Hieu. Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, pages 177–180.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Hubert Jin. 2004. Arabic treebank: Part 3 v 1.0. In *Linguistic Data Consortium*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, Volume 29, No 1*, pages 19–51.

Arfath Pasha, Mohamed Al-Badrashiny, Ramy Kholy Ahmed El Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC 2014, Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Tachicart Ridouane and Bouzoubaa Karim. 2014. A hybrid approach to translate moroccan arabic dialect. In *SITA'14, 9th International Conference on Intelligent Systems*.

Wael. Salloum and Nizar. Habash. 2012. Elissa: A dialectal to standard arabic machine translation system. In *Coling'2012, 24th International Conference on Computational Linguistics*, pages 385–392.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT 13, pages 348–358.

Hassan. Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *AMTA'2010, 9th Conf. of the Association for Machine Translation in the Americas*.

Inguna Skadiņa, Ahmet Aker, Voula Giouli, Dan Tufis, Robert Gaizauskas, Madara Mieriņa, and Nikos Mastropavlos. 2010. A Collection of Comparable Corpora for Under-resourced Languages. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 161–168.

Andreas Stolcke. 2002. Srilm – an Extensible Language Modeling Toolkit. In *ICSLP*, pages 901–904, Denver, USA.

Rabih Zbib, Erika Malchiodi, Devlin Jacob, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT 12, pages 49–59.