HAL
open science

Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation

Salima Harrat, Karima Meftouh, Kamel Smaïli

To cite this version:

Salima Harrat, Karima Meftouh, Kamel Smaïli. Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation. Computación y sistemas, inPress, 23 (3), pp.979-989. 10.13053/cys-23-3-3267 . hal-02274533

HAL Id: hal-02274533

https://hal.science/hal-02274533v1

Submitted on 30 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation

Salima Harrat[1], Karima Meftouh[2], Kamel Smaïli[3]

[1] École Normale de Bouzaréah, Algiers
Algeria

[2] Badji Mokhtar University-Annaba
Algeria

[3] Campus scientifique LORIA
France

slmhrrt@gmail.com , karima.meftouh@univ-annaba.dz, kamel.smaili@loria.fr

Abstract. This research deals with resources creation for under-resourced languages. We try to adapt existing resources for other resourced-languages to process less-resourced ones. We focus on Arabic dialects of the Maghreb, namely Algerian, Moroccan and Tunisian. We first adapt a well-known statistical word segmenter to segment Algerian dialect texts written in both Arabic and Latin scripts. We demonstrate that unsupervised morphological segmentation could be applied to Arabic dialects regardless of used script. Next, we use this kind of segmentation to improve statistical machine translation scores between the tree Maghrebi dialects and French. We use a parallel multidialectal corpus that includes six Arabic dialects in addition to MSA and French. We achieved interesting results. Regards to word segmentation, the rate of correctly segmented words reached 70% for those written in Latin script and 79% for those written in Arabic script. For machine translation, the unsupervised morphological segmentation helped to decrease out-of-vocabulary words rates by a minimum of 35%.

Keywords. Arabic dialects , Morphological segmentation , Machine translation.

1 Introduction

The linguistic situation of the Arab world is characterized by the diglossia phenomenon, which is the co-existence of two variants of the same language. A standard language (standard Arabic) used in formal speeches, newspapers, education, etc. Arabic dialects which are informal languages used in everyday conversations. Natural language processing of Arabic language does not take into account a large wide of these dialects, which lack NLP resources until today. These vernaculars are considered as under-resourced languages. Compared to other under-resourced languages, these dialects bring particular challenges because of their oral nature. They were not written until the advance of Internet and mobile telephony. They have no standard rules that normalize their transcription therefore a word is written in different forms which are all acceptable. Nowadays, Arabic dialects are widely used in social networks. They are written in Arabic and Latin script.[1] Also, They are written sometimes with a mixture of letters and numbers. Arab people exploit the similarity between some Arabic letters and numbers to write the dialect, for example similarity between 3 and ع, 7 and ح and 9 and ق. These dialects are variants of Arabic language; they are different from it and also they differ from each other. Maghrebi dialects are different from Middle-east dialects.

[1] Arabic dialect written in Latin script is called in recent research: Arabizi, Arabish or Romanized Arabic.

Also, in the same Arab country several dialects exist. In addition, these dialects are evolving, new dialectal words appear every day and are adopted without academic validation. Even, Arabic dialects are influenced by other foreign languages such as French, Spanish, Turkish and Berber (for Maghrebi dialects). This influence generates the code-switching phenomenon, a dialectal sentence could include words from two or three languages. It is common to alternate between dialect, standard Arabic, French or English in the same conversation.

In this paper, we focus on Maghrebi Arabic dialects. We use a data-driven approach for word segmentation of Algerian dialect texts. We adapt Morfessor (a well-known statistical word segmenter); we present for the first time to our knowledge, a segmenter that considers dialectal texts written in both Arab and Latin scripts. In addition, we investigate the impact of word segmentation on machine translation performance. We use for this purpose statistical machine translation (SMT) from the three Maghrebi dialects (Algerian, Moroccan and Tunisian) to French. To do this, we present a new version of a parallel corpus previously created and containing six Arabic dialects besides MSA. This new version includes for the first time a French text. The rest of this article is organized as follows: we first describe briefly Arabic dialects, particularly, Magherbi ones (Section 2). Then, we provide an overview of related work on Arabic dialect morphological segmentation. Section 3 is dedicated to the adaptation of Morfessor for unsupervised and semi-supervised segmentation of Algerian dialect texts. In Section4, we investigate the impact of morphological segmentation on statistical machine translation from Maghrebi dialects to French. Section 5 concludes this paper by pointing future directions of our work.

2 Arabic dialects, Focus on the Three Main Maghrebi Dialects

Arabic dialects (vernaculars or colloquial Arabic) are considered as one of three variants of Arabic language, which includes also classical Arabic and Modern standard Arabic (MSA).[2] Arabic dialects are a spoken form of Arabic, used in everyday conversations, they are different from one Arabic country to another. They are influenced by both local tongues and foreign languages such as Spanish, French, Italian and English. In terms of classification, Arabic dialects are distinguished regards to the East-west dichotomy[23]: (a) Middle-east dialects which include spoken Arabic of Arab Gulf countries and Yemen, Iraqi dialect, Levantine dialect (Syria, Lebanon, Palestine and Jordan), besides Egyptian and Sudanese dialects. (b) Maghrebi dialects which include the dialects of Algeria, Tunisia, Morocco, Libya and Mauritania. As already mentioned before, we focus in this paper on the three main Maghrebi dialects: Algerian, Tunisian and Moroccan, one can raise the question: why only these dialects? The reason is that these dialects are the only ones for which we have relatively available resources. In addition, the three Maghrebi countries share a lot of social, cultural, religious and linguistic similarities. Regarding the linguistic side, in the three countries, the Berber is the oldest language which has coexisted until now with the Arabic language bring to the region with Islamic conquest. The Algerian, Tunisian and Moroccan dialects are mutually intelligible, speakers of the three countries can readily understand each other. They share a lot of common features, even though they are different from each other. More extensive comparative details of the three dialects could be found in [20].

3 Morphological Segmentation of Algerian Dialect Texts

3.1 Related work

Many efforts have been dedicated to build morphological segmenters for Arabic dialects texts. There are for this issue two main approaches; building segmenters from scratch [17, 5] or adapting MSA ones to take into account dialectal features. Several studies adopted this last

[2] Classical Arabic is the Arabic of the Quran and the ancient literature of Arabian peninsula while MSA is a modern form of classical Arabic.

approach. Authors of [35] used the well known morphological analyzer BAMA[38] by extending its affixes tables to Levantine and Egyptian dialects. In the same way, BAMA was adapted to deal with Algerian dialect [19], the authors rebuilt affixes and stems tables. They kept MSA entries that apply also to Algerian dialect and integrated purely dialectal entries. Similarly, in [4], Al-Khalil morphological segmenter [8] has been adapted by enriching its affixes dictionary with a list of affixes belonging to four Arabic dialects. Likewise, the authors of the work described in [15] converted an Egyptian lexicon (ECAL, Egyptian Colloquial Arabic Lexicon) into a representation similar to the SAMA [14] dictionary (Standard Modern Arabic Analyzer). It should be noted that all these segmenters are dedicated to texts written in Arabic script.

3.2 Motivation

Our goal is to segment Algerian dialects texts regardless of their script. To that end, we adopt an adaptive approach. However, we do not adapt a MSA morphological segmenter but rather a morphological segmenter based on probabilistic machine learning methods. The following reasons justify this choice:

— As mentioned above, dialectal texts are written in different forms with no standard orthography. They are written with Arabic and Latin script and sometimes with numbers instead of letters. This lack of writing rules is a challenging issue for morphological segmentation. Hence, data-driven approaches seem to be the most appropriate solution for this task.

— Non-standard spelling of dialects texts makes rule-based approach difficult to consider.

— Because of the evolving nature of dialectal vocabulary, new words appear and are rapidly spread in speakers' community. Data-driven methods could easily take into account this words and their inflected forms.

3.3 Morfessor

In this respect, we opted for Morfessor, a well-known morphological segmenter suitable for languages with complex morphology like Finnish and Turkish. It has been integrated into different NLP applications like speech recognition [24, 30, 13, 36], machine translation [41, 27, 29, 9, 33] and speech retrieval [7, 39].
Morfessor [10, 11] is a set of statistical methods for segmenting words based on the Minimum Description Length principle. It learns morphemes from data in an unsupervised manner. The level of segmentation is tuned by adjusting the weight α between the cost of encoding the lexicon (the parameters Θ) and the cost of encoding the training data (D) part in the cost function:

$$L(\Theta, D_w) = -\log P(\Theta) - \alpha \log P(D_w|\Theta). \quad (1)$$

An interesting version of Merfossor is that described in [26] where a semi-supervised training approach is used. The above function cost is summarized as follows:

$$L(\Theta, D_w) = -\log P(\Theta) - \alpha \log P(D_w|\Theta) - \beta \log P(A|\Theta). \quad (2)$$

where A is the annotated training data, and α and β in this order, are the weights of the unannotated and annotated data training. In the context of this work, we use the Morfessor 2.0 implementation [40].

3.4 Data Description

In order to train Morfessor, we used textual corpora recently created in the context of processing Algerian dialect. Below, we give an overview of each corpus.

— The comparable corpus CALYOU
CALYOU[3][1] is an Algerian dialect comparable corpus of Youtube comments. It was collected by querying Youtube with key-words related to current Algerian events. The corpus includes comments written with Arabic script aligned to ones written with Latin script. This alignment is got by using word embeddings. We give in

[3] Comparable spoken ALgerian extracted from YOUtube

Tables 1 and 2 respectively, statistics about this corpus and some comments examples written with Arabic and Latin scripts including even numbers.

Table 1. CALYOU corpus statistics

#Comments(K)	#Words(M)	#Distinct words(K)
853	12.7	88

Table 2. Examples of CALYOU comments with mapping between Arabic letters and numbers

Dialectal comment	Meaning
فور بزاف	It is very good
خاوتي ديروا ابوني.	Brothers! subscribe
علاخطش نحبو بلادنا برافو برافو	Because we love our country bravo
3andk l7a9 (3=ع , 9=ق)	You are right
ya3tik asaha madame (3=ع)	Thank you madam
sa7bi a9ra l'histoire ta3 bladek (7=ح)	My friend, you must learn the history of your country
taktal badahk 5ouya (5=خ)	You are very funny my brother

— Algerian text of PADIC (ALG-PADIC)
PADIC[4][28] is multidialectal Arabic corpus including Algerian, Tunisian, Morrocan, Syrian and Palestinian in addition to MSA. We use for the purpose of this work, the Algerian side of this corpus (see statistics in Table 3).

Table 3. ALG-PADIC corpus statistics

#Sentences(K)	#Words(K)	#Distincts words(K)
6.4	40.75	9.15

[4]Parallel Arabic Dialect Corpus, downloadable on `http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/`

3.5 Experimentation

The experiments were carried out using the corpora described above. We experimented unsupervised segmentation trained with CALYOU corpus. Then, we conducted a semi-supervised training of Morfessor with annotated data provided from the ALG-PADIC corpus.
For evaluation purpose, we randomly extracted two datasets of 200 CALYOU comments written in Arabic and Latin scripts with respectively, 1730 and 1609 words. The two test datasets has been segmented by hand.

3.5.1 Unsupervised Morphological Segmentation

We trained Morfessor with CALYOU corpus. In order to tune the weight α that controls segments lengths (a low α favors small construction lexicons, while a high value favors longer constructions), we made several experiments starting with the default value ($\alpha = 1$). The figure 1 retraces the results in terms of percentage of correctly segmented words written in Latin and Arabic scripts according to the different values of α.

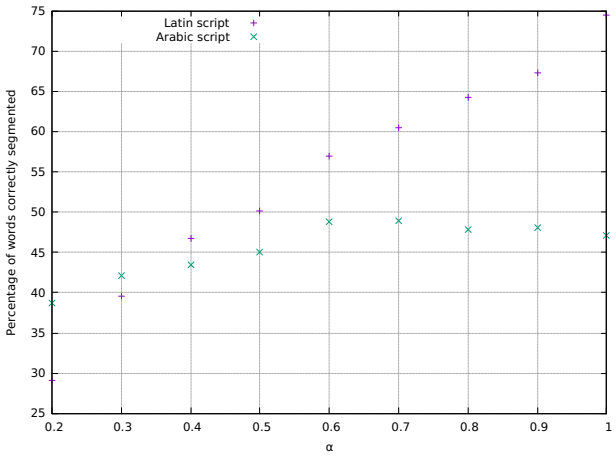

Fig. 1. Percentages of correctly segmented words using unsupervised method with different values of α

It shows that 74.46% of words written with Latin script in the test set are correctly segmented for the default value of α which proved to be the best of all the ones we tested. Indeed, the segmentation

Table 4. Examples of words written with Latin script correctly segmented using unsupervised method

Segments	Word	Segmentation	Meaning
Conjunction+demonstrative pronoun.	whada	w+hada	And this one
Conjunction+noun	wrajel	w+rajel	And a man
Definition article+noun	l3sal	l+ 3sal	The honey
Function word+pronoun	3andna	3and+na	We have
Preposition+pronoun	mnhoum	mn+houm	From them
Subject-prefix+verb	yadkhol	ya+dkhol	He enters
Verb+suffix-subject	kbarty	kbar+ty	You have grown

Table 5. Examples of words written with Latin script partially segmented using unsupervised method

Word	Partial segmentation	Valid segmentation	Meaning
mayjouzch	ma+yjouz+ch	ma+y+jouz+ch	He does not pass
yaatik	ya+atik	ya+ati+k	He gives you
may3arfakch	may+3arfak+ch	ma+y+3arf+ak+ch	He does not know you

Table 6. Examples of words written with Arabic script correctly segmented using unsupervised method

Segments	Word	Segmentation	Meaning
Conjunction + personal pronoun	وانت	و+انت	And you
Definition article+noun	الدار	ال+دار	The house
Function word+pronoun	عندي	عند+ي	I have
Verb+subject suffix+ object suffix	سبقونا	سبق+و+نا	They have surpassed us
Preposition+noun	بدراهم	ب+دراهم	With money
Preposition+noun+ suffix pronoun	لبنتك	ل+بنت+ك	For your daughter

takes into account several morphological features like function words inflection. We give in Table 4 some examples of valid segmentations provided by the test set.

Furthermore, for invalid segmentations, we noticed that in most cases Morfessor could identify some segments of the word even though he could not identify all the segments. For example, the circumfix negation affixes are often distinguished (see examples in Table 5).

For Words written with Arabic script, unsupervised segmentation performs worse than words with Latin script. The best-recorded percentages are got for an α value of 0.7 and do not reach 50%. However, even for Arabic script words, Morfessor could identify correctly some segments of a word although the whole segmentation is not valid. Tables 6 and 7 show some illustrative examples.

3.5.2 Semi-supervised Morphological Segmentation

In this experiment, we performed tests with semi-supervised segmentation. Unfortunately, we did it for dialect texts written with Arabic script only, since annotation data for Latin script are not available for us. Indeed, we used the ALG-padic

Table 7. Examples of words written with Arabic script partially segmented using unsupervised method

Word	Partial segmentation	Valid segmentation	Meaning
تعيشي	ت+عيشي	ت+عيش+ي	You live
جربتو	جرب+تو	جرب+ت+و	I tried it
وماجاش	و+ماجا+ش	و+ما+جا+ش	And he did not come

corpus for annotation. We have segmented it using the morphological analyzer [19] described earlier.[5]

Morfessor is thus trained with CALYOU corpus and ALG-padic annotated corpus. It should be noted that in addition to the α parameter already described, Morfessor uses another parameter β that controls the contribution of the annotation data in the segmentation operation. We first started by using the segmentation with the default values, then we experimented different values of α and β. We show in Figure 2 the best achieved results in terms of percentages of correctly segmented words.

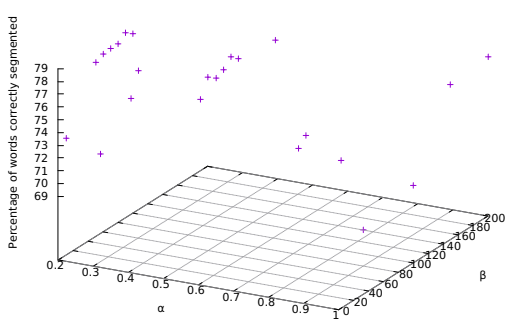

Fig. 2. Percentages of correctly segmented words using semi-supervised method with different values of α and β

Semi-supervised segmentation shows promising results regards to the size of the annotated corpus. Indeed, the best percentage of correctly segmented word reached 78.55%. According to the test sample, Semi-supervised segmentation could take into account many dialectal morphological features. Most agglutinated forms that the unsupervised segmentation failed to segment where correctly analyzed by the semi-supervised analysis. We report in Table 8 some examples.

Furthermore, despite its ability to segment agglutinated forms correctly, we noticed that even with semi-supervised analysis, the negation forms is difficult to segment for words written in Arabic script. Morfessor failed to identify all word segments. Some examples are reported in Table 9.

We also found that for some words, semi-supervised analysis tends to over-segment. In Table 10 are given some examples of these cases.

Morfessor segmentation seems to be an interesting direction for segmenting dialectal Arabic words, in view of the fact that it can be used for texts in Arabic and Latin scripts. Moreover, transcribing dialect by introducing numbers is not problematic with Morfessor. Words written with numbers are segmented as well as words including only letters. Verb conjugation and noun declension are taken into account as illustrated above in the various examples. In addition, through the different segmentations that we analyzed, the agglutinative forms of the dialects (more complicated than MSA) are for the most part parsed.

4 Impact of Morphological Segmentation on SMT of Maghrebi Dialects to French

Word segmentation is an important step in many NLP tasks related to Arabic. Many work show that it improves performance of NLP applications like part-of-speech tagging[12,

[5]We remind that this Analyzer supports the Arabic script only.

Table 8. Examples of words correctly segmented with semi-supervised method

Segments	Word	Segmentation	Meaning
Conjunction+noun+plu. suffix+pronoun suff.	ووليداتك	و+وليد+ات+ك	And your children
Conj.+subj.pref.+verb+subj. suff.+obj. suff.	ونعاودوها	و+ن+عاود+و+ها	And we will repeat it
Subj. pref.+ verb+ object suff.	يخلصهم	ي+خلص+هم	He pays them
Function word+pronoun suff.	علينا	علي+نا	On us
Def. article+noun+plural suff.	الـمومنين	ال+مومن+ين	The believers

Table 9. Examples of words written with Arabic script partially segmented with semi-supervised method

Word	Partial segmentation	Valid segmentation	Meaning
ونتمنالك	و+ن+تمن+ال+ك	و+ن+تمنا+ل+ك	I wish you
ماتـحكيليش	مات+حكي+ل+ي+ش	ما+ت+حكي+ل+ي+ش	Do not tell me
ماتحبوش	مات+حبو+ش	ما+ت+حب+و+ش	You do not like it

Table 10. Example of over-segmented words with semi-supervised method

Word	Over Segment.	Valid Segment.	Meaning
سخون	سخ+و+ن	سخون	Hot
معسله	مع+سل+ه	معسل+ه	Honeydew
برافو	ب+راف+و	برافو	Bravo

16, 34] and machine translation [18, 2, 3, 34]. In this respect, we attempt to measure the impact of unsupervised segmentation on machine translation performance in the context of translating between Arabic Maghrebi dialects (Algerian, Tunisian and Moroccan) and French. It should be noted that, most research efforts in this area concern English. For more details, the reader is referred to [21] where a comprehensive survey on Arabic dialects machine translation is presented.

4.1 Settings

We use a phrase-based statistical machine translation [25], with Giza++[31] for alignment and KenLM [22] to compute ngram language models. We also use an unsupervised segmentation with Morfessor. We choose unsupervised segmentation because annotation data are not available for Tunisian and Moroccan dialects, they are available only for Algerian dialect.

4.2 Data Description

1. Parallel corpus of Maghrebi dialects and French
   We use the three Maghrebi dialect texts of PADIC corpus (Algerian, Moroccan and Tunisian) and for the first time, a parallel French text translated from the standard Arabic side of PADIC (see Table 11).

2. Monolingual corpora
   For unsupervised training of Morfessor, three monolingual dialectal corpora are used. A brief description of these corpora is given below (with some statistics in Table 12):

Table 11. The parallel corpus statistics

Corpus	#Words (K)	#Distinct word (K)
Algerian	40.75	9.15
Moroccan	42.58	9.70
Tunisian	38.96	10.04
French	62.02	7.91

- Arabic script part of CALYOU used earlier.
- A Tunisian corpus of facebook comments [6] collected during the period of Arab spring events (we used only comments written with Arabic script).
- A Moroccan corpus of texts[37] collected from different sources (web sites, plays and records of everyday conversations).

Table 12. Statistics of dialectal monolingual corpora used for unsupervised segmentation

Corpus	#Words(K)	#Distinct words (K)
Algerian	412.93	191.17
Moroccan	349.30	62.87
Tunisian	131.46	32.25

Also we used a French monolingual corpus which we downloaded from OPUS[6] web site to train French language models.

4.3 Experimentation

We trained all the machine translation systems on 5.9K parallel sentences. We allocated 0.1K and 0.4K sentences for tuning and evaluation, respectively. The baseline SMT systems (SMT-no-segmentation) are trained on unsegmented data for the three dialects (source language being the dialect and the target language French). Next, we segmented data for training, tuning and evaluating SMT systems. Regards to the size of monolingual dialectal corpora used for learning Morfessor, we conducted two types of experiments. In the first one, data of the three SMT systems (SMT+seg(32K)) are segmented by learning Morfessor with datasets of 32K distinct words for each dialect (32K is the size of Tunisian corpus, the smallest monolingual corpus). In the second experiment, SMT systems (SMT+seg(62K)) data were segmented by learning Morfessor with datasets of 62K distinct words. This experiment concerns only Algerian and Moroccan dialects because we have no more data for Tunisian dialect. We evaluated all SMT systems described in terms of BLEU[32] metric.

Table 13 shows results. We notice that for Tunisian dialect-to-French translation, SMT system that uses Morfessor segmentation outperforms the system that does not use segmentation by 1.25 BLEU points. For Algerian-to-French and Moroccan-to-French SMT systems whose data were segmented by Morfessor learned with datasets of 32K words, BLEU scores decrease by 0.61 and 0.93 points respectively. However, when Morfessor is learned with more dialectal data (62K words), BLEU score of Algerian-to-French increases by 0.41 compared to the baseline system score. For Moroccan-to-French translation, the BLEU score of SMT+seg(62K) system (segmentation learned on a dataset of 62K words) outperforms the SMT+seg(32K) system (segmentation learned on a dataset of 32K words). But, the baseline system remains the best. Furthermore, Morfessor segmentation decreases significantly OOV rates. Indeed, OOV rates of Algerian-to-French and Tunisian-to-French SMT systems trained on segmented data decrease by almost 50%. For Moroccan-to-French SMT system, Morfessor does not improve BLEU scores as seen, but it decreases the OOV rates by at least 34%.

5 Conclusion

We have adopted an unsupervised and semi-supervised approach to segment Algerian dialect texts written in Arabic and Latin scripts. This work was accomplished by using Morfessor. The results are encouraging. Indeed, most morphological features of Algerian dialects are taken into

[6] http://opus.nlpl.eu/

Table 13. BLEU scores and OOV rates of Maghrebi dialects to French SMT according to different training data of Morfessor

SMT systems	Algerian		Moroccan		Tunisian	
	BLEU	OOV%	BLEU	OOV%	BLEU	OOV%
SMT-no-segmentation	6.90	24.7	9.01	23.5	7.43	28.3
SMT+seg(32K)	6.29	16.2	8.08	15.1	8.68	14.5
SMT+seg(62K)	7.31	12.6	8.84	15.5	-	-

account. The semi-supervised segmentation applied only for text written in Arabic script achieved the best results. We further evaluated the impact of unsupervised segmentation of Maghrebi dialect texts on SMT systems that translate from the three Maghrebi dialects to French. For the first time, we introduced a French text to PADIC corpus. This text was used to train the target side of the SMT systems. The unsupervised segmentation has improved BLEU scores especially for Tunisian-to-French and Algerian-to-French SMT systems. Moreover, the OOV rates decrease by nearly 50% for these two SMT systems and by more than 34% for the Moroccan-to-French SMT system. In the future, we would like to use an iterative process to create annotation data for Algerian dialect texts in Latin script. We will use unsupervised segmentation to segment the words, then valid outputs will be used to create annotation data. This will allows us to consider semi-supervised segmentation for these texts. In the same way, we will enrich the annotated data in Arabic script. Finally, the new version of PADIC will be made available to the scientific community.

References

1. **Abidi, K., Menacer, M.-A., & Smaili, K. (2017).** Calyou: A comparable spoken Algerian corpus harvested from youtube. *18th Annual Conference of the International Communication Association (Interspeech)*.

2. **Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., & Vogel, S. (2014).** Unsupervised word segmentation improves dialectal Arabic to English machine translation. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pp. 207–216.

3. **Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016).** First result on Arabic neural machine translation. *arXiv preprint arXiv:1606.02680*.

4. **Almeman, K. & Lee, M. (2012).** Towards developing a multi-dialect morphological analyser for Arabic. *4th International Conference on Arabic Language Processing*, pp. 19–25.

5. **Altantawy, M., Habash, N., & Rambow, O. (2011).** Fast yet rich morphological analysis. *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, Association for Computational Linguistics, pp. 116–124.

6. **Ameur, H. & Jamoussi, S. (2013).** Dynamic construction of dictionaries for sentiment classification. *2013 IEEE 13th International Conference on Data Mining Workshops*, IEEE, pp. 896–903.

7. **Arisoy, E., Can, D., Parlak, S., Sak, H., & Saraclar, M. (2009).** Turkish broadcast news transcription and retrieval. *Trans. Audio, Speech and Lang. Proc.*, Vol. 17, No. 5, pp. 874–883.

8. **Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O., & Shoul, M. (2011).** Alkhalil morpho sys: A morphosyntactic analysis system for Arabic texts. *Proceedings of 7th International Computing Conference in Arab ACIT*.

9. **Clifton, A. & Sarkar, A. (2011).** Combining morpheme-based machine translation with post-processing morpheme prediction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, pp. 32–42.

10. **Creutz, M. & Lagus, K. (2002).** Unsupervised discovery of morphemes. *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, Association for Computational Linguistics, pp. 21–30.

11. **Creutz, M. & Lagus, K. (2007).** Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, Vol. 4, No. 1, pp. 3.

12. **Diab, M., Hacioglu, K., & Jurafsky, D. (2004).** Automatic tagging of Arabic text: From raw text to base phrase chunks. *Proceedings of HLT-NAACL 2004: Short papers*, Association for Computational Linguistics, pp. 149–152.

13. **Gelas, H., Besacier, L., & Pellegrino, F. c. (2012).** Developments of Swahili resources for an automatic speech recognition system. *Spoken Language Technologies for Under-Resourced Languages*.

14. **Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., & Buckwalter, T. (2009).** Standard Arabic morphological analyzer (SAMA) version 3.1. *Linguistic Data Consortium LDC2009E73*.

15. **Habash, N., Eskander, R., & Hawwari, A. (2012).** Morphological analyzer for Egyptian Arabic. *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON*, Association for Computational Linguistics, pp. 1–9.

16. **Habash, N. & Rambow, O. (2005).** Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 573–580.

17. **Habash, N. & Rambow, O. (2006).** Magead: A morphological analyzer and generator for the Arabic dialects. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 681–688.

18. **Habash, N. & Sadat, F. (2006).** Arabic preprocessing schemes for statistical machine translation. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, Association for Computational Linguistics, pp. 49–52.

19. **Harrat, S., Meftouh, K., Abbas, M., & Smaili, K. (2014).** Building resources for Algerian Arabic dialects. *Proceedings of Interspeech*, pp. 2123–2127.

20. **Harrat, S., Meftouh, K., & Smaïli, K. (2018).** Maghrebi Arabic dialect processing: an overview. *Journal of International Science and General Applications*, Vol. 1.

21. **Harrat, S., Meftouh, K., & Smaili, K. (2019).** Machine translation for Arabic dialects (survey). *Information Processing & Management*, Vol. 56, No. 2, pp. 262 – 273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

22. **Heafield, K. (2011).** Kenlm: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 187–197.

23. **Hetzron, R. (1997).** *The Semitic Languages*. Routledge language family descriptions. Routledge.

24. **Hirsimaki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., & Pylkkönen, J. (2006).** Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*, Vol. 20, No. 4, pp. 515–541.

25. **Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007).** Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, pp. 177–180.

26. **Kohonen, O., Virpioja, S., & Lagus, K. (2010).** Semi-supervised learning of concatenative morphology. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, Association for Computational Linguistics, pp. 78–86.

27. **Luong, M.-T., Nakov, P., & Kan, M.-Y. (2010).** A hybrid morpheme-word representation for machine translation of morphologically rich languages. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 148–157.

28. **Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaili, K. (2015).** Machine translation

experiments on PADIC: A Parallel Arabic DIalect Corpus. *Proceedings PaCLiC 29th Asia Conference on Language, Information and Computation*, pp. 26–34.

29. **Mermer, C. (2010).** Unsupervised search for the optimal segmentation for statistical machine translation. *Proceedings of the ACL 2010 Student Research Workshop*, Association for Computational Linguistics, pp. 31–36.

30. **Mihajlik, P., Tuske, Z., Tarjan, B., Nemeth, B., & Fegyo, T. (2010).** Improved recognition of spontaneous Hungarian speech; morphological and acoustic modeling techniques for a less resourced task. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1588–1600.

31. **Och, F. J. & Ney, H. (2003).** A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, volume 29, number 1*, pp. 19–51.

32. **Papineni, K. & al. (2001).** Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual of the Association for Computational linguistics*, Philadelphia, USA, pp. 311–318.

33. **Popović, M. (2011).** Morphemes and POS tags for n-gram based evaluation metrics. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 104–107.

34. **Sajjad, H., Dalvi, F., Durrani, N., Abdelali, A., Belinkov, Y., & Vogel, S. (2017).** Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. *arXiv preprint arXiv:1709.00616*.

35. **Salloum, W. & Habash, N. (2011).** Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, Association for Computational Linguistics, pp. 10–21.

36. **Smit, P., Leinonen, J., Jokinen, K., Kurimo, M., et al. (2016).** Automatic speech recognition for Northern Sami with comparison to other Uralic languages. *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages*, The Research Group on Artificial Intelligence (RGAI).

37. **Tachicart, R., Bouzoubaa, K., Aouragh, S. L., & Jaafa, H. (2018).** Automatic identification of Moroccan colloquial Arabic. *Arabic Language Processing: From Theory to Practice*, Springer International Publishing, pp. 201–214.

38. **Tim, B. (2002).** Buckwalter Arabic morphological analyzer version 1.0. *Linguistic Data Consortium LDC2002L49*.

39. **Turunen, V. T. & Kurimo, M. (2008).** Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Trans. Speech Lang. Process.*, Vol. 8, No. 1, pp. 1:1–1:25.

40. **Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013).** Morfessor 2.0: Python implementation and extensions for Morfessor Baseline.

41. **Virpioja, S., Väyrynen, J. J., Creutz, M., & Sadeniemi, M. (2007).** Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, Vol. 2007, pp. 491–498.

Article received on dd/mm/yyyy; accepted on dd/mm/yyyy.
Corresponding author is Salima Harrat.