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

Salima Harrat1, Karima Meftouh2, Kamel Smaili3

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 script1. 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 خ, 7 and ح and 9 and ق. These dialects are variants of Arabic language;

---

1Arabic dialect written in Latin script is called in recent research: Arabizi, Arabish or Romanized Arabic.
they are different from it and also they differ from each other.

Maghrebi dialects are different from Middle-east dialects. 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, Maghrebi 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 Section 4, 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). 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].

\[2\] Classical Arabic is the Arabic of the Quran and the ancient literature of Arabian peninsula while MSA is a modern form of classical Arabic.
3 Morphological Segmentation of Arabic Dialect Texts

3.1 Related Work

Many efforts have been dedicated to build morphological segmenters for Arabic dialect 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 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 $\alpha$ between the cost of encoding the lexicon (the parameters $\Theta$) and the cost of encoding the training data (D) part in the cost function:

$$L(\Theta, D_w) = -\log P(\Theta) - \alpha \log P(D_w|\Theta).$$  \hspace{1cm} (1)

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

$$L(\Theta, D_w) = -\log P(\Theta) - \alpha \log P(D_w|\Theta) - \beta \log P(A|\Theta),$$  \hspace{1cm} (2)

where $A$ is the annotated training data, and $\alpha$ and $\beta$ 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$^3$[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 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

<table>
<thead>
<tr>
<th>#Comments(K)</th>
<th>#Words(M)</th>
<th>#Distinct words(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>853</td>
<td>12.7</td>
<td>88</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dialectal comment</th>
<th>Meaning</th>
</tr>
</thead>
</table>
| قور يزار
خاونتي ديرو ابوينى
لاخطش قمبا بلادنا يزاروف
برافو | It is very good
Brothers! subscribe
Because we love our country
bravo |
| 3andk 17a9 غ=3 ق=9 | You are right
ya3tik asaha madame (3=ع) | Thank you madam
| sa7bi a9ra lhistoire ta3 bladek (7=ح) | My friend, you must learn the history of your country |
| taktal badakh 5ouya (5=ح) | You are very funny my brother |

— Algerian text of PADIC (ALG-PADIC)
PADIC[^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).

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 $\alpha$ that controls segments lengths (a low $\alpha$ 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 $\alpha$.

![Fig. 1. Percentages of correctly segmented words using unsupervised method with different values of $\alpha$](image)

It shows that 74.46% of words written with Latin script in the test set are correctly segmented for the

[^28]: Parallel Arabic Dialect Corpus, downloadable on http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/
default value of $\alpha$ which proved to be the best of all the ones we tested.

Indeed, the segmentation 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 $\alpha$ 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 corpus for annotation. We have segmented it using the morphological analyzer [19] described earlier.

Morfessor is thus trained with CALYOU corpus and ALG-padic annotated corpus. It should be noted that in addition to the $\alpha$ parameter already described, Morfessor uses another parameter $\beta$ 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 $\alpha$ and $\beta$. We show in Figure 2 the best achieved results in terms of percentages of correctly segmented words.

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.

---

5We remind that this Analyzer supports the Arabic script only.
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, 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.
Table 8. Examples of words correctly segmented with semi-supervised method

<table>
<thead>
<tr>
<th>Segments</th>
<th>Word</th>
<th>Segmentation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjunction+noun+plur. suffix+pronoun suff.</td>
<td>و+وليدهات+ك</td>
<td>½K@ YJ Ëðð Ëð Æ</td>
<td>And your children</td>
</tr>
<tr>
<td>Conj.+subj.pref.+verb+subj. suff.+obj. suff.</td>
<td>و+عن+هاوود+ها</td>
<td>¼ + H@ + YJ Ëð Æ</td>
<td>And we will repeat it</td>
</tr>
<tr>
<td>Subj. pref.+verb+object suff.</td>
<td>ي+خلص+هم</td>
<td>ي+خلص+هم</td>
<td>He pays them</td>
</tr>
<tr>
<td>Function word+pronoun suff.</td>
<td>علنيا</td>
<td>علنيا</td>
<td>On us</td>
</tr>
<tr>
<td>Def. article+noun+plural suff.</td>
<td>ال+مومنين</td>
<td>ال+مومنين</td>
<td>The believers</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Partial segmentation</th>
<th>Valid segmentation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>و+تمثالك</td>
<td>و+تمثالك</td>
<td>مات+حكي+بل+ي+ش</td>
<td>I wish you</td>
</tr>
<tr>
<td>ماتحكي+ش</td>
<td>ماتحكي+ش</td>
<td>مات+حبوءش</td>
<td>Do not tell me</td>
</tr>
<tr>
<td>مات+حبوءش</td>
<td>مات+حبوءش</td>
<td>مات+حبوءش</td>
<td>You do not like it</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Over Segment</th>
<th>Valid Segment</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>حطون</td>
<td>حطون</td>
<td>Hot</td>
<td></td>
</tr>
<tr>
<td>مع+مل+ه</td>
<td>مع+مل+ه</td>
<td>Honeydew</td>
<td></td>
</tr>
<tr>
<td>ب+زااف+و</td>
<td>ب+زااف+و</td>
<td>Bravo</td>
<td></td>
</tr>
</tbody>
</table>

French text translated from the standard Arabic side of PADIC (see Table 12).

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 13):

- 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).

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

\(^6\)http://opus.nlpl.eu/
Table 11. BLEU scores and OOV rates of Maghrebi dialects to French SMT according to different training data of Morfessor

<table>
<thead>
<tr>
<th>SMT systems</th>
<th>Algerian</th>
<th></th>
<th>Moroccan</th>
<th></th>
<th>Tunisian</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>OOV%</td>
<td>BLEU</td>
<td>OOV%</td>
<td>BLEU</td>
<td>OOV%</td>
</tr>
<tr>
<td>SMT - no-segmentation</td>
<td>6.90</td>
<td>24.7</td>
<td>9.01</td>
<td>23.5</td>
<td>7.43</td>
<td>28.3</td>
</tr>
<tr>
<td>SMT + seg(32K)</td>
<td>6.29</td>
<td>16.2</td>
<td>8.08</td>
<td>15.1</td>
<td>8.68</td>
<td>14.5</td>
</tr>
<tr>
<td>SMT + seg(62K)</td>
<td>7.31</td>
<td>12.6</td>
<td>8.84</td>
<td>15.5</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 12. The parallel corpus statistics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words (K)</th>
<th>#Distinct word (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algerian</td>
<td>40.75</td>
<td>9.15</td>
</tr>
<tr>
<td>Moroccan</td>
<td>42.58</td>
<td>9.70</td>
</tr>
<tr>
<td>Tunisian</td>
<td>38.96</td>
<td>10.04</td>
</tr>
<tr>
<td>French</td>
<td>62.02</td>
<td>7.91</td>
</tr>
</tbody>
</table>

Table 13. Statistics of dialectal monolingual corpora used for unsupervised segmentation

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words (K)</th>
<th>#Distinct words (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algerian</td>
<td>412.93</td>
<td>191.17</td>
</tr>
<tr>
<td>Moroccan</td>
<td>349.30</td>
<td>62.87</td>
</tr>
<tr>
<td>Tunisian</td>
<td>131.46</td>
<td>32.25</td>
</tr>
</tbody>
</table>

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 11 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 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 allow 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


Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON, Association for Computational Linguistics, pp. 1–9.


the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 104–107.


Article received on 24/01/2019; accepted on 04/03/2019.
Corresponding author is Salima Harrat.