# Parallel resources for Tunisian Arabic dialect translation

**Saméh Kchaou**
University of Sfax, Tunisia
samehkchaou4@gmail.com

**Rahma Boujelbane**
University of Sfax, Tunisia
rahmaboujelban@gmail.com

**Lamia Hadrich Belguith**
University of Sfax, Tunisia
Lamia.belguith@gmail.com

## Abstract

The difficulty of processing dialects is clearly observed in the high cost of building representative corpus, in particular for machine translation. Indeed, all machine translation systems require a huge amount and good management of training data, which represents a challenge in a low-resource setting such as the Tunisian Arabic dialect. In this paper, we present a data augmentation technique to create a parallel corpus for Tunisian Arabic dialect written in social media and standard Arabic in order to build a **M**achine **T**ranslation (MT) model. The created corpus was used to build a sentence-based translation model. This model reached a BLEU score of 15.03% on a test set, while it was limited to 13.27% utilizing the corpus without augmentation.

## 1 Introduction

Nowadays, the use of the dialect in social networks is becoming more and more important. Besides dialects emerge as the language of informal communication on the web (forums, blogs, emails and social media). Thus, there is a shift from a purely oral language to a written language, without established standardization or standard spelling. It is also obvious that the dialect becomes intensively employed by internet users in social networks to share their thoughts and opinions. This kind of writing is sometimes incompressible even by persons speaking the same dialect, especially Arabic written using the latin script. Consequently, it is necessary to transform the text from its informal type to a formal type in order to be understandable and suitable for **N**atural **L**anguage **P**rocessing (NLP) tools. In this work, we create parallel resources in order to build a MT model able to transform the **T**unisian **D**ialect (TD), written in social networks, into the **M**odern **S**tandard **A**rabic (MSA). This represents a motivating task for both academic and industrial fields. Indeed, switching to a standard language facilitates communication between people overall the word. It also makes automatic opinion analysis easier using well-equipped language tools. For instance, businesses operating on the global market rely on social media campaigns to promote their brand and products, creating engaging and catchy content that can be exploited by users. However, to our knowledge, until now there is no works that have dealt with the translation of TD, especially the one that exists in social networks.

We introduce, in this study, an augmentation technique in order to build useful resources to train an efficient TD-MSA MT model. For this, we collected a significant number of parallel written resources to start with a total word count of about 75k. Then, we segmented the MSA sentences and we generated partial parallel sentences using back translation based on a bilingual dictionary created from the collected corpus. Finally, in order to trained the efficiency of the proposed resources, we test a statistical MT model on different configurations of the created corpus. Obviously, the translation performance is improved by augmenting the size of TD-MSA parallel data sets. The remainder of this paper is structured as follows: In section 2, we review related work on data augmentation. In section 3, we describe the process of building parallel resources. Section 4 depict the different steps of building a statistical translation model. We discuss in section 5 the performance of the resulting model.

.

*Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 200–206
Barcelona, Spain (Online), December 12, 2020

## 2 Related Work

Machine translation for low-resource language is a well-known task in the NLP community. Several works focused on it to form efficient MT models. In order to train these models, different configurations of parallel resources have been proposed. There are those who have proven that sentence segmentation associated with the technique of back translation is a very suitable method for constructing parallel resources for low resourced languages. For example, (Jinyi and Tadahiro, 2019) applied this technique to generate pseudo-parallel pairs of Japanese-Chinese sentences by dividing the long parallel sentence pair of the corpus into parallel partial sentences pairs and using back-translation from the target partial sentence. The pair of long parallel sentences was divided into segments at the level of punctuation marks such as ",", ";", ":". This method improved the performance of translation. (Alina et al., 2018) proposed also a back translation method for low resource languages. The authors used an iterative back-translation of monolingual low resource data with the models trained on the transliterated high-resource data and utilized the resulting parallel corpus to train the final models. Similarly, another method was suggested by (Rico et al., 2015) to generate an English-German and Turkish-English parallel corpus using the back-translating of monolingual target data into the source language in order to obtain pseudo-source sentences. The back-translation was based on a statistical translation model. The size of the corpus was increased by the pairs of pseudo-source sentences and the original target sentences. Otherwise, (Guillaume et al., 2018) proposed two models : a neural model and a sentence-based model in order to generate automatically a parallel data using the iterative back-translation and denoising the effect of the language models formed on the target side. (Marzieh et al., 2017) generated new pairs of sentences containing rare words in new synthetically-created contexts. Researchers used language models trained on a large amount of German monolingual data to produce new English-German and German-English sentence pairs. (Fei et al., 2019) presented a soft contextual data augmentation method in order to build neural machine translation. The method consists in replacing a randomly-chosen word in a sentence with a soft distributional representation provided by a language model. The translation results proved the effectiveness of this method. A new efficient data augmentation approach to improve German-English translation performance was presented in (Li et al., 2020). Authors proposed a diversified data augmentation strategy that does not use extra monolingual data. They trained a forward and backward model to translate the training data.

## 3 TD-MSA parallel resources

Parallel corpus aligned at the sentence level are essential resources to build MT systems. One of our goals is to collect the largest possible set of parallel TD-MSA sentences. We collected in the first step, the existing free parallel corpus used in the state of the art. Secondly, we translated manually a corpus scraped from social networks.

### 3.1 Existing resources

**Parallel Arabic DIalectal Corpus (PADIC) :** It is a parallel corpus combining Maghreb dialects (Algerian, Tunisian and Moroccan), Levant dialects (Palestinian and Syrian) and the MSA (Karima et al., 2015). It was built from two sub-corpora : the first was created by recording different conversions from everyday life of the Annaba dialect, and the second sub-corpora was formed by the recordings which correspond to films and television programs expressed in the dialect of Algeria [1]. Then, the transcribed data was translated into the Moroccan dialect. Afterwards, these two corpora of the Annaba dialect and the Algeria dialect were translated into MSA. MSA was subsequently used as a pivot language to obtain the other Tunisia, Syrian and Palestinian dialects. This corpus was utilized to develop several statistical translation systems between these dialects and the MSA. We extracted a sub-corpus containing 6.4K TD/MSA sentences per language. Sentence length varied between 1 and 29 words on the MSA side, and between 1 and 27 on the TD side.

[1] https://sites.google.com/site/torjmanepnr/6-corpus

**Multi Arabic Dialect Applications and Resources (MADAR) :** It is the second TD-MSA parallel free available corpus (Bouamor et al., 2018) built within the framework of the MADAR project [2]. This fine-grained corpus consists of several English sentences selected from the **C**orpus **B**asic **T**raveling **E**xpression (BTEC) (Takezawa et al., 2006) and translated by native speakers into 25 different Arabic dialects plus MSA. It contains 1.8k sentences per language. In this study, we use only the pair of languages (TD-MSA). The long MSA sentences include 31 words inside MSA and 28 inside TD.

**Tunisian CONSTitution (TD-CONST) :** After the Tunisian revolution, the constitution written in MSA was translated into Tunisian Arabic dialect to be more understandable by the Tunisian community. We aligned the TD version with that of the MSA. This gave us a corpus containing about 500 parallel laws, in which the length of MSA sentences is between 3 and 37 words, and that of Tunisian sentences is between 3 and 27 words. This corpus was not used in previous researches dealing with machine translation.

### 3.2 Social media corpus

In order to adapt the existing corpus to the textual content of the social networks, we scraped 900 Tunisian **COM**ments (TD-COM) ) from Facebook. This corpus was then translated into MSA by a native speaker. We got TD-MSA parallel sentences composed of 31 MSA words as maximum words per sentence, and 30 words in TD side.
Thus, through the different sources, we collected 9.7K parallel sentences. Table 1 shows more statistics on this corpus.

| **Corpus** | **#Lines** | **#TD words** | **#MSA words** | **#Distinct TD Words** | **#Distinct MSA Words** |
|---|---|---|---|---|---|
| PADIC | 6.4k | 38.6k | 43.6k | 10k | 9.64k |
| MADAR | 1.8k | 9.8k | 11.8k | 3.8k | 4.1k |
| TD-CONST | 600 | 8.3k | 7.7k | 2.6k | 2.4k |
| TD-COM | 900 | 11.4k | 11.1k | 5.8k | 5.3k |
| All corpus | 9.7k | 68k | 74k | 18.17k | 17.53k |

Table 1: Statistics of the collected corpus

### 3.3 TD-MSA dictionary

We created a bilingual TD-MSA dictionary from the alignment of the collected parallel corpus. Indeed, we employed the statistical machine translation toolkit GIZA ++[3] (Och and Hermann, 2003) to form a HMM word alignment model with which an automatically align words was generated between the source and target sides of the dataset. Then, we checked manually the obtained lexicon in order to remove errors. The resulting lexicon contains 44k TD-MSA entries.

### 3.4 TD – MSA lexical differences

The dialect language, by virtue of its utilitarian nature, evolves much more rapidly than the standard language. We can now consider these two forms of a single language as two distinct languages although they are clearly related. According to (Boukadida, 2008), Arabic dialect differs from MSA on all levels of linguistic representation. We distinguished through the collected corpus, the lexical difference that exist between TD and MSA. Indeed, there are sentences that do not change when migrating to the MSA. For instance, this sentence : صبَاح الخير *.spA.h Alxyr "Good morning"* is used in both TD and MSA. Other ones change partially lexicons when migrating to MSA. For example, the TD sentence عندي دفوَار نهَار الخميس *'ndy dfwAr nhAr Alxmys* becomes in MSA عندي اختبَار يوم الخميس *'ndy*

[2] https://sites.google.com/view/madar-shared-task/home
[3] https://github.com/moses-smt/giza-pp

*axtbAr ywm Alxmys ”I have a test on Thursday”.*
There is also sentences that change totally the lexicon. For example, the MSA sentence المكَان جَميل جدًا *mkAn jamyl jdA ”The place is very nice”* becomes in TD محلَاه البلَاصة تهَبِل *ma.hlAh AlblA.sT thabel.*
Thus in order to explore the degree of similarity between the TD and MSA in the collected corpus, we generated the intersection set between the two vocabularies. Figure 2 shows the lexical similarity between these two languages in the different corpora. There is on average 30% of common words between TD and MSA, which proves the hypothesis that, although TD and MSA are clearly related, they can be considered as two different languages; hence the need for a translation process.

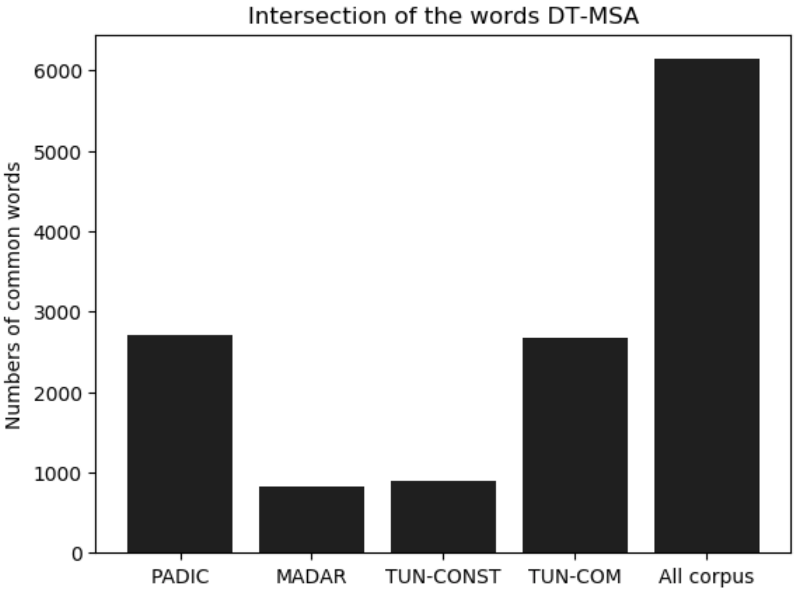

Figure 1: Intersection of the TD-MSA words.

### 3.5 Corpus pretreatment

**Text segmentation :** A part of the TD-CONST corpus contains a parallel text without any punctuation mark to delimit the sentences. So, based on the meaning of the sentence and referring to a native speaker, we proposed a segmentation to this corpus. This treatment allowed obtaining a set of 600 sentences while it was 500 parallels texts before the segmentation.
We then proposed a segmentation for all the MSA sentences of the corpus. In fact, as there is no TD segmentation tools or standard syntax rules, we segmented MSA sentences based on Arabic stop words and punctuation marks.
**Back translation :** The segmentation process resulted in a loss of pairwise alignment in many sentences. Figure 2 shows an example of misalignment of partial sentences after the segmentation of the MSA sentence. To remedy this problem, we opted for the back translation using the created MSA-DT dictionary in order to re-translate each MSA target partial sentences into corresponding TD sentences.

## 4 Statistical machine translation model

In order to test the efficiency of the resource-created method in the context of machine translation, we adapted a simple traditional sentence-based **S**tatistical **M**achine **T**ranslation (SMT) model. Indeed, deep learning-based methods have made significant progress in recent years in the context of machine translation. We opted for the same methodology of (Och and Hermann, 2003). The latter has shown its effectiveness in several similar works as it is the case in (Guillaume et al., 2018).

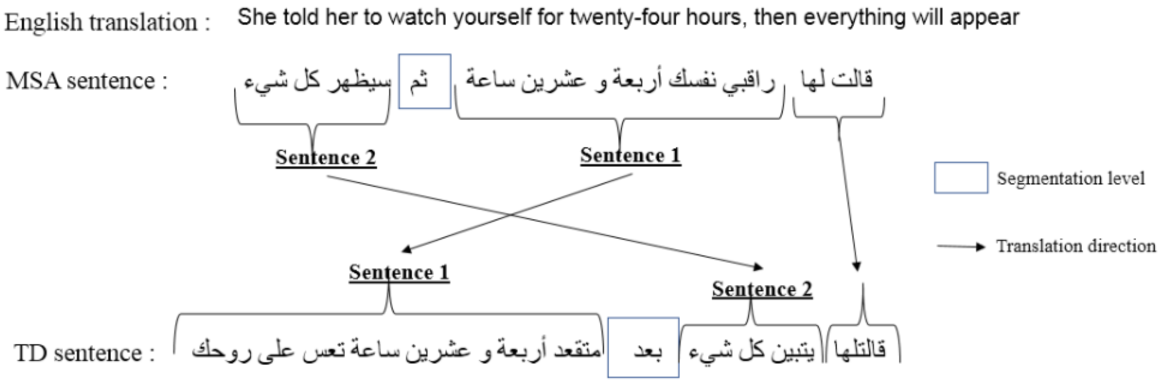

Figure 2: Example of TD-MSA sentence pair alignment.

### 4.1 Word alignment

The basic step of applying the proposed model is word-to-word translations. For this reason, we started with word alignment using the TD-MSA bi-text corpus. We applied GIZA ++ defined in the previous section in order to align the source and target sentence. Giza ++ takes a pair of sentences as input: source sentence matches its target sentence and returns the word alignment of the target sentence for the given source language. Then, sentence tables are created to store the probability that a certain n-gram in the source language is mapped to another n-gram in the target language. This step returns all possible sentences associated with a source sentence.

### 4.2 Comments translation probabilities

Sentence transition probabilities consists in assigning a probability for each sentence in the training corpus obtained in the previous phase by calculating the relative occurrences of the target sentence c for a given source sentence s for both directions. The sentence translation probability is given by the following equation.

$$\phi(\tilde{s}|\tilde{c}) = \frac{count(\tilde{s}|\tilde{s})}{\sum_{\tilde{c}_i} count(\tilde{s}|\tilde{c}_i)} \tag{1}$$

In addition to translation probabilities, we trained language models on source and target languages. The language model allowed improving the translation quality by sampling the structure and reorganization of words. We used the IRSTLM [4] toolkit to train the language models. IRSTLM: the IRST Language Modeling Toolkit is utilized to calculate statistical n-gram language models. It is integrated into the moses SMT decoder.

The final probabilities of the sentence were computed by combining the translation probability from the language model with the probability of translating the sentence from the source-target alignment matrix.

### 4.3 Phrase translation

In order to generate a sentence translation, the model used stack-based decoding to obtain the best translation from the sentence translation probabilities provided by training and to from the language model. The experiment was performed to translate TD sentences into MSA sentences.

## 5 Experimental results & discussion

In this section, we present an analysis of the SMT model applied on the created resources. We divided these resources into two sets: train set and test set. As we want to build a translation model for TD comments in social networks, we considered 500 parallel sentences among the 900 sentences of the TD-COM corpus as a test set. The remaining 400 sentences of the TD-COM corpus were used in the

[4] https://sourceforge.net/projects/irstlm/

train set. We carried out three experiments to analyze the utilized data. We increased the train corpus gradually in order to test the effect of the data augmentation on the applied model, as shown in Table 2.

**Experiment 1 :** We started by training the model with the available parallel data. We employed firstly TD-MSA data from the MADAR and PADIC corpus. We got a BLEU score of 13.07% on the test set. Then, we added the TD-CONST data set collected from the Tunisian constitution. This enrichment by the constitution data degraded the result of translation to 12%. This score can be explained by the nature of vocabulary utilized in the Tunisian constitution corpus which is different from that of social networks. Afterwards, we integrated the data into 400 sentences from the TD-COM corpus. The BLEU score was improved to 13.27% on the same Test set. Following this extension, we trained a model on the following configuration of corpus : MADAR, PADIC and TD-COM. The resulting model achieved an improvement in the BLEU score that rose to 13.67%. We can conclude, from this experiment, that the increase in corpus is preferably done with a corpus of the same domain or close domain.

**Experiment 2 :** Based on the results of experiment 1, we considered the corpus composed by the sub-corpora of MADAR, PADIC and TD-COM. We segmented all long sentences of this corpus. The aim of this experiment is to test the influence of the sentence size on the quality of translation. On the segmented corpus version, resulting from the back translation, we have learned again the SMT model. By this textual representation, the performance of the model has degraded. The model achieved a BLUE score of 11.58%. The dictionary-based translation led to the loss of meaning of some sentences and, consequently, reduced the score obtained in this experiment.

**Experiment 3 :** In this experiment, we combined, in the training set, the pairs of the segmented sentences with the pairs of the original sentences in the corpus. This improved the performance of the model which achieved a BLUE score of 15.03%.
Table 2 summarizes the variation of the BLUE score of the different models learned in the three experiments. All models were tested on the same test set. Thus, we can conclude from these different experiments that the injection of a parallel corpus of the target domain is very beneficial to improve translation. Besides, augmenting the corpus by a segmented version can further enhance the obtained results.

| **Training set** | **#Lines** | **BLEU (%)** |
|---|---|---|
| PADIC + MADAR | 8.2k | 13.07 |
| PADIC + MADAR + TD-CONST | 8.8k | 12 |
| PADIC + MADAR + TD-COM | 8.6k | 13.67 |
| PADIC + MADAR + TD-COM + TD-CONST | 9.2k | 13.27 |
| (PADIC + MADAR + TD-COM) segmented | 32K | 11.58 |
| (PADIC + MADAR + TD-COM) augmented | 41k | **15.03** |

Table 2: TD-MSA translation BLEU score.

## 6 Conclusion

We presented, in this paper, a process of creating parallel TD-MSA resources by exploiting firstly the available resources. Then, we enriched the data by building a parallel corpus scraped from social media. Afterwards, we proposed a corpus augmentation technique based on segmenting the well-endowed language and back translation. In order to test the effectiveness of the constructed resources, we trained a SMT model using different configurations of collected corpus and we tested it on a test corpus taken from social media. Experimental results show that the ”augmented” configuration achieved better result of translation with a BLEU score of 15.03%.
We aim in future work, to further increase the size of the corpus and include other types of comments

like those that contain code switching. We also aim to test the efficiency of the neuronal approach to translate the written comments in social networks.

## References

Karakanta Alina, Dehdari Jon, and van Genabith Josef. 2018. Neural machine translation for low-resource languages without parallel corpora. In *Machine Translation 2018, 32*, page 167–189.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, may.

Nahed Boukadida. 2008. Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (etude longitudinale).

Gao Fei, Zhu Jinhua, Wu Lijun, Xia Yingce, Qin Tao, Cheng Xueqi, Zhou Wengang, and Liu Tie-Yan. 2019. Soft contextual data augmentation for neural machine translation. In *Association for Computational Linguistics*, Florence, Italy.

Lample Guillaume, Ott Myle, Conneau Alexis, and Denoyer Ludovic. 2018. Phrase-based neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 5039–5049.

Zhang Jinyi and Matsumoto Tadahiro. 2019. Corpus augmentation by sentence segmentation for low-resource neural machine translation. *CoRR*, abs/1905.08945.

Meftouh Karima, Harrat Salima, Jamoussi S, Abbas M, and Smaïli Kamel. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of 29th Paclic Asia Conference on Language, Information and Computation.*, pages 26–34.

Y Li, X. Li, Y. Yang, and R. Dong. 2020. A diverse data augmentation strategy for low-resource neural machine translation. In *Information 2020, 11, 255.*, pages 2078–2489.

Fadaee Marzieh, Bisazza Arianna, and Monz Christof. 2017. Data augmentation for low-resource neural machine translation. In *Proc. 55th Annual Meeting of the Assoc. for Computational Linguistics.*, page 567–573, Vancouver, Canada.

Franz Josef Och and Ney Hermann. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, page 19–51.

Sennrich Rico, Haddow Barry, and Birch Alexandra. 2015. Improving neural machine translation models with monolingual data. In *Actes de la 54e réunion annuelle de l'Association for Computational Linguistics.*, page 86–96, Berlin, Allemagne.

Toshiyuki Takezawa, Kikui Genichiro, Mizushima Masahide, and Sumita Eiichiro. 2006. Multilingual spoken language corpus development for communication research. In *Chinese Spoken Language Processing*, pages 781–791.