Improving Machine Translation of Arabic Dialects through Multi-Task Learning

Youness Moukafih, Nada Sbihi, Mounir Ghogho, Kamel Smaïli

Improving Machine Translation of Arabic Dialects through Multi-Task Learning

Youness Moukafih[1,2], Nada Sbihi[1], Mounir Ghogho[1], and Kamel Smaili[2]

[1] TICLab, College of Engineering and Architecture, Université Internationale de Rabat, Morocco
{youness.moukafih,mounir.ghogho,nada.sbihi}@uir.ac.ma
[2] LORIA/INRIA-Lorraine 615 rue du Jardin Botanique, BP 101, F-54600 Villers-16s-Nancy, France
{youness.moukafih,kamel.smaili}@loria.fr

Abstract. Neural Machine Translation (NMT) systems have been shown to perform impressively on many language pairs compared to Statistical Machine Translation (SMT). However, these systems are data-intensive, which is problematic for the majority of language pairs, and especially for low-resource languages. In this work, we address this issue in the case of certain Arabic dialects, those variants of Modern Standard Arabic (MSA) that are spelling non-standard, morphologically rich, and yet resource-poor variants. Here, we have experimented with several multitasking learning strategies to take advantage of the relationships between these dialects. Despite the simplicity of this idea, empirical results show that several multitasking learning strategies are capable of achieving remarkable performance compared to statistical machine translation. For instance, we obtained the BLUE scores for the Algerian $\rightarrow$ Modern-Standard-Arabic and the Moroccan $\rightarrow$ Palestinian of 35.06 and 27.55, respectively, while the scores obtained with a statistical method are 15.1 and 18.91 respectively. We show that on 42 machine translation experiments, and despite the use of a small corpus, multitasking learning achieves better performance than statistical machine translation in 88% of cases.

Keywords: Neural Network · Machine Translation · Multitask Learning · Low-resource Languages · Arabic dialects.

1 Introduction

Arabic dialects are morphologically rich vernaculars, just like Modern Standard Arabic (MSA), this leads to some challenges in automatic language processing in general and in machine translation in particular. In the Arab world, at least two languages coexist in a single country, one of which is MSA. This phenomenon of coexistence of languages used by the same linguistic community is known, in linguistics, under the name of diglossia [11]. The modern standard Arabic is unique, has a standard orthography and is used in formal settings such as broadcast news, religious speeches, governmental documents, and other printed material, while Arabic dialects are several, they are considered as the mother tongues of the population that depend on the their born regions.

These latter have no standard written form and are used mostly for verbal communication [17]. Arabic dialects can be clustered in two groups: the Mashriqi (eastern) dialects group, which includes dialects of Arabian Peninsula, Mesopotamia, Levant, Egypt and Sudan, and the Maghrebi (western) dialects group, which are characterized by a high level of code-switching in the daily communication. While mutual intelligibility within each group is high, it is not so between the two groups, e.g., Moroccan and Palestinian. This reason could justify the development of Machine translation among these dialects. [5] studied the linguistic variation among Arabic dialects to that among Romance languages, indicating the need for machine translation between these dialects. However, while most machine translation systems have been conducted for rich-resource language pairs, [1, 20], only very limited number of works tackled the issue of Dialect-to-dialect pairs due to the problem of parallel data sparsity [2, 10, 16, 18]. Arabic dialects are under-resourced languages, in addition, they face two other issues: *morphological richness*, *orthographic ambiguity*. None of these issues are unique to Arabic dialect, but their combination makes DA processing particularly challenging.

Morphological Richness : Arabic (MSA and Arabic dialects) is a morphologically complex language which includes rich inflectional morphology and a high number of clitics. For instance, the Moroccan dialect word " وَغَايْكَتْبُوهَا " correspond to the English phrase "and they will write it". This phenomenon leads to a higher number of unique words compared to English, which has the consequence to increase the number of entries in the vocabulary and therefore necessitates a bigger parallel corpus for the training that are not available for Arabic dialects.

Orthographic Ambiguity: The Arabic script uses optional diacritical marks to represent short vowels and other phonological information that are useful for removing the ambiguity [13]. For instance, the word كتبت (without diacritics) could correspond to several other words, among them: كَتَبَتْ (*she wrote*), كَبتُ (*I wrote*. Arab speakers do not generally have a problem with reading undiacritized text, they use the context in order to remove the ambiguity depending on the position of the no-vowled word. However, for computers or beginners in Arabic, this task is very challenging.

Many attempts were proposed to handle the issue of translating these dialects to each other or from or into MSA. However, most of these system are based on either rule-based approach or a statistical machine translation approach. This last one used to dominate MT research for decades. For instance, in [16, 18], the authors developed statistical machine translation systems between several Arabic dialects using Parallel Arabic DIalect Corpus (PADIC) [3].
Another work, based on neural network approach, proposed by [2], in which the authors proposed a multi-task learning method for translating dialectal Arabic to MSA by leveraging a pivot Language (English). However, due to the aforementioned prob-

[3] https://smart.loria.fr/corpora/

lems of Arabic dialects, adding English or any other structured language do not help the model due to the gap between the different data distributions of these different languages.
In this paper, we adopt a different approach where we leverage the closeness between Arabic dialects and perform simultaneous translations of multiple dialect pairs using a neural multi-task learning framework. This alleviates the issue of parallel data sparsity. To the best of our knowledge, our work is the first to use the translations of multiple Arabic dialect pairs as related tasks in a multi-task learning setup. Our approach outperformed previous statistical machine translation and achieved state-of-the-art result on 88% of the translation directions of the pairs of languages of PADIC corpus.

The remainder of this paper is structured as follows. In the next section, we discuss related work. Section 3 provides a detailed description of PADIC dataset. Section 4 describes the proposed method. In section 5, we present the results of the proposed machine translation approach using several language pairs, and compare these results with those of other learning strategies. Finally, conclusions are drawn in Section 6.

2 Related Work

Sequence to sequence models are the common choice for machine translation systems in most language pairs. However, these models are rarely used in Arabic dialects due to scarcity of parallel corpora. A lot of work on the translation from Arabic dialects to MSA, based on rule-based methods, has been carried out. For example, [21] proposed a rule-based approach that relies on language modeling to translate from Moroccan dialect to MSA by adapting many tools such as Alkhalil morphological analyzer [4], which was initially developed for MSA. In [9], the authors used a rule-based method to improve Egyptian-English translation by identifying a mapping from Egyptian dialect to MSA to reduce the out-of-vocabulary rate. [12] proposed a machine translation system for both TUN to MSA and MSA to TUN based on deep morphological representations of roots and patterns' features. The system reached about 80% recall in the TUN to MSA direction and 84% recall in the opposite direction. All the methods used in the above-mentioned papers focused on a rue-based approach which requires enormous amount and linguistic resources.

[3] tackled the challenge of translating from Arabic dialects (Levantine dialects and Maghrebi dialects) to MSA by using a neural machine translation system. In that work, the authors used a multi-task learning paradigm by sharing one decoder between two target languages (MSA and English) and each source language has an encoder (the sources languages were Arabic dialects and MSA). Another interesting work was presented in [2] which proposed a unified multitask neural machine translation model where an encoder is shared between two tasks, the first task being Arabic Dialect to MSA translation and the second task being segment-level Part-Of-Speech (POS) tagging. The model achieved a definite improvement of the translation performance, and a good performance on the test set for the POS tagging task.

[16, 18] presented the PADIC dataset which consists of parallel sentences in Levantine dialects (Syrian and Palestinian), Maghrebi dialects (Moroccan, two dialects from Algeria and Tunisian) and MSA. The authors proposed a statistical machine translation method by employing different smoothing techniques for the language model to translate not only from Arabic dialects to MSA but also between all language pairs within PADIC dataset. The obtained results were relatively good given the size of the training corpus, especially for similar languages.

3 DATASET

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. Arabic dialects are among those languages for which the parallel corpora are scarce.
In this paper, we use The Parallel Arabic DIalect Corpus (PADIC) [16]. The corpus (containing 273k words) has been built from scratch because there are no standard resources. Indeed, Arabic dialects are only used in daily oral communication and social networks and not for formal writing. PADIC contains six dialects from both the Maghreb and the Middle-East as well as MSA. The dialects are: Annaba's dialect (ANB) and Algiers's dialect (ALG) which are Algerian dialects, Moroccan dialect (MAR), Sfax's dialect (TUN) used in the south of Tunisia, Syrian dialect (SYR) and Palestinian dialect (PAL) which are spoken in Damascus and Gaza respectively.

4 METHODOLOGY

4.1 Sequence-to-sequence learning

Here, we describe briefly the underlying framework, called Encoder-Decoder architecture. The encoder-decoder with recurrent neural networks has two components:
An encoder reads the input sentence, a sequence of words (x_1,x_2,...,x_T), where x_t is the t^{th} word, and produces a context vector c_i which encodes sentence information with strong focus on the parts surrounding the i^{th} word, as shown in Equation 1 :

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{1}$$

where $h_t = f(x_t, h_{t-1})$, with:

- $h_t \in \mathcal{R}^d$ being the hidden state at time t, and d is the the dimension of the hidden state vector.
- f being a nonlinear activation function (LSTM or GRU).

– α_{ij} being the so-called energy, computed by:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{2}$$

where

$$e_{ij} = a(s_{i-1}, h_j) \tag{3}$$

with a being a feed-forward neural network trained jointly with all the other components of the model that scores how well the inputs around the position j and the output at position i match, and s_{i-1} being the previous hidden state of the decoder.

On the other hand, a decoder takes as inputs the context vector c_i and all the previously predicted words (y_1,...,y_{i-1}) trained to predict the next word y_i as in Equation 4

$$p(\mathbf{y}) = \prod_{i=1}^{T_y} p(y_i | \{y_1, ..., y_{i-1}\}, c_i) \tag{4}$$

where $\mathbf{y} = \{y_1,...,y_{T_y}\}$. With an RNN, each conditional probability is modeled as follows:

$$p(y_i | \{y_1, ..., y_{i-1}\}, c_i) = g(y_{i-1}, s_i, c_i) \tag{5}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_i, and s_i is an RNN hidden state for time i, computed as in Equation 6.

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{6}$$

4.2 Multi-task sequence-to-sequence learning

Multi-Task Learning (MTL) has been used successfully in many domains such as natural language processing [7], speech recognition [8], and computer vision [15].
The use of multi-task learning in this work is motivated by the relatively small size of PADIC corpus and also by the idea that learning one encoder across multiple language pairs jointly may result in a better generalization because of the similarities between the languages considered here.

Learning multiple tasks simultaneously can be applied in different ways such as periodic task alternations training with a ratio for each task based on the size of the task's data-set. In this work, we used the simplest approach to train these multiple tasks jointly by taking a mini-batch of data per task for each training iteration and update the model's parameters for every mini-batch. Figure 1 illustrate an example where the model takes first as input MSA-to-MAR mini-batch data; the encoder encodes the MSA sentences and the decoder takes the encoded vector and produces the Moroccan translation sentences, then the model takes the second mini-batch (ALG-to-MAR) and so on.

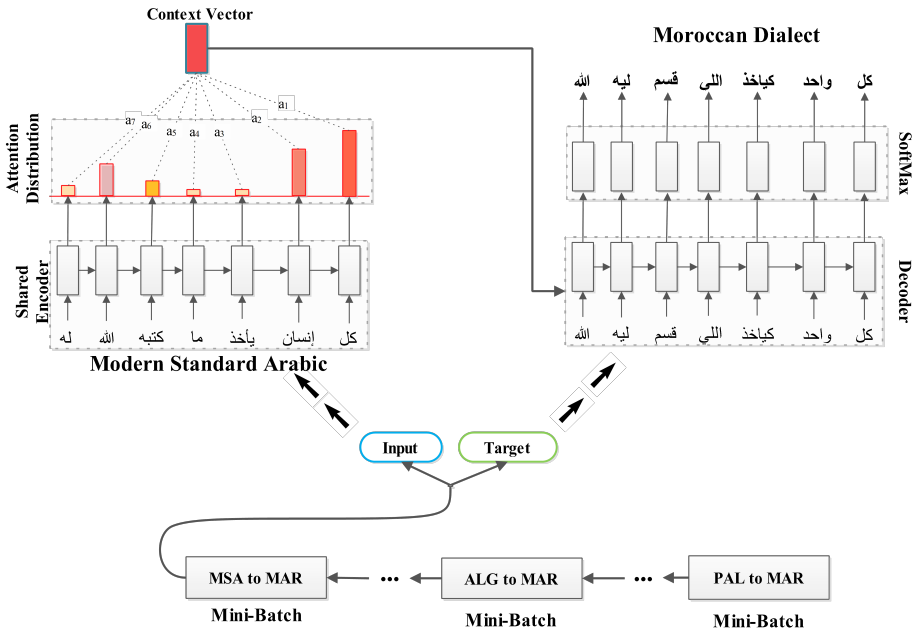

Fig. 1. Illustration of the architecture of the used multitask sequence-to-sequence learning with an attention mechanism. The model takes a mini-batch of data from each language pair.

4.3 Model training

In this work, the translation model is composed of one encoder, shared among several translation directions, and a decoder for each target language. The main objective is to develop a model which is able to map an input sentence from a chosen language ℓ_k in $\{\ell_1, \ell_2, ...\ell_L\}$ to a target language ℓ_t, where L is the number of input languages we use for our encoder. Mathematically, given a sentence $S_k \sim \mathrm{D}_{\ell_k}$ of m words $x = (x_1, x_2, ..., x_m)$ an encoder $E(\ x,\ \ell_k,\ \theta_E)$ produces a representation vector $h_k \in \mathrm{R}^n$ where θ_E are weights shared across all input languages and n is the dimension of the hidden states, then a decoder $D(\ h_k,\ \ell_t,\ \theta_D)$ takes as input the vector representation and generates an output sentence $y = (y_1, y_2, ..., y_q)$. The objective function is defined as follows:

$$\mathcal{L}(\theta_E, \theta_D, Z, \ell_{in}, \ell_o) = E_{x \sim \mathcal{D}_{\ell_{in}}, y \sim \mathcal{D}_{\ell_o}} [\Delta(y, \hat{y})] \tag{7}$$

where Z is the set of word embeddings, ℓ_{in} is the input languages, ℓ_{ou} is the output language, $\hat{y}$ is the predicted sentence, and Δ is the sum of token-level cross-entropy losses.

5 EXPERIMENTAL RESULTS & DISCUSSION

5.1 Experiment settings

We use Pytorch [19] library to implement all our experiments. We set the size of the embedding vectors to 256 for all languages. The embedding vectors are initialized randomly. Each mini-batch used in the training consists of 32 sentences randomly selected from each translation direction with equal ratios for each language pair. In each mini-batch, we trained the encoder-decoder model using GRUs [6], each having 256 hidden units, to minimize the sum of token-level cross-entropy losses provided in Equation 7 using Adam optimizer [14] with a learning rate of $\alpha = 0.003$.

Performance of the Many-to-One multi-task learning approach We recall that we opted for one shared encoder that encodes several dialects and one decoder for the target language. In other words, any target language among the languages used in the encoding step could be decoded by one decoder. We will refer to this multi-task learning strategy as Many-to-One (M-2-O).

Table 4 shows the results of this Multi-task learning model and we compare them to those obtained by the statistical machine translation approach presented in [16, 18]. The results are given in terms of BLEU score measured on the same corpus of 500 unseen parallel sentences. In bold, we reported the best performances. It is shown that in the majority of cases (88%), our approach achieves better results than the statistical approach; the gain in performance is significant for many language pairs. For the M-2-O approach, the lowest and the highest BLEU scores are respectively 22.75 for the pair Moroccan-Syrian, and 35.96 for the pair Algiers-Palestinian, while the lowest and the highest BLEU scores for the statistical model are respectively 7.29 for the pair Algiers-Syrian and 61.06 for the pair Annaba-Algiers. The latter result is due to the fact that the corresponding dialects are from the same country; 60% of words are shared between the two dialects in accordance to the study presented in [16, 18]. The method that we propose here has not achieved this high score because the aim is to learn a general-purpose sentence representations across all translation directions. But, except few cases, as indicated our method is better than the statistical approach in 88% of cases. The proposed learning model is shown to have the potential to learn sentence representations across all language pairs and produce non-trivial translations, which confirms the effectiveness and the robustness of the approach.

It is evident from Table 4 that the results achieved by our proposed multi-task learning approach are significantly better than those obtained with the statistical model [16]. This could be attributed to the fact the architecture of the proposed multi-task learning model is, thanks to its sophistication, capable of capturing more relationships between the source and the target sentences, and also by the fact that this model benefits from more data since all the entire corpus has been used to train one neural network model, whereas in the statistical model, only data corresponding to a pair of languages is used in the training. In addition, our proposed system address one of the weakness in conventional

	Target													
	MSA		**ALG**		**ANB**		**TUN**		**PAL**		**SYR**		**MAR**	
Source	M-2-O	SMT	M-2-O	SMT	M-2-O	SMT	M-2-O	SMT	M-2-O	SMT	M-2-O	SMT	M-2-O	SMT
MSA	———	———	**27.71**	13.55	**27.27**	12.54	**27.96**	20.03	30.79	**42.46**	**26.55**	21.38	**28.45**	20.02
ALG	**35.06**	15.1	———	———	30.41	**61.06**	**32.39**	09.67	**35.96**	10.61	**32.64**	07.29	**33.23**	10.22
ANB	**30.81**	14.44	27.55	**67.31**	———	———	**29.56**	09.08	**31.63**	10.12	**31.82**	07.52	**33.38**	10.00
TUN	25.42	**25.99**	**24.61**	09.89	**25.04**	09.34	———	———	**27.88**	22.55	**25.95**	13.05	**25.29**	14.37
PAL	34.50	**40.48**	**33.61**	11.28	**33.39**	09.53	**34.64**	17.93	———	———	**32.46**	23.29	**33.86**	16.08
SYR	**27.85**	24.14	**26.94**	07.57	**25.05**	07.50	**24.98**	13.67	**27.88**	26.60	———	———	**26.73**	09.93
MAR	**26.18**	24.93	**24.01**	10.13	**23.33**	10.16	**23.82**	14.68	**27.55**	18.91	**22.75**	09.68	———	———

Table 1. Comparison of the performance of the M-2-0 Multi-task learning model and the statistical model of [16]

NMT systems which is their inability to correctly translate very rare words. Table 2 summarises the improvement in performance achieved by the proposed model over the statistical one. The first column indicates the average BLEU corresponding to the translation of any language to a specific target language. For instance, the first line corresponds to the average performance achieved from any dialect to the the Moroccan dialect. In our experiments, the best performance is achieved for pairs of dialects where the target language is the Moroccan dialect (an improvement of 124.49%).

	Many-to-One	**Statistical**	**Rate (%)**
Any-to-MAR	30.15	13,43	124,49
Any-to-ALG	27,40	19,95	37,34
Any-to-ANB	27,41	18,35	49,37
Any-to-TUN	28,89	14,17	103,88
Any-to-PAL	30,27	21,87	38,40
Any-to-SYR	28,69	13,70	109,41
Any-to-MSA	29,97	24,18	23,94

Table 2. A summary of the results of different machine translation methods for several pairs of dialects

Performance of single-task neural network We have carried out other experiments based on a simple sequence-to sequence neural network in order to determine the impact of using neural network machine translation when using a small training corpus. The single task model (S-task), used in this paper, has one encoder and one decoder and it is trained only in one translation direction. For instance, from Moroccan dialect to Algerian dialect the encoder will take as input the Moroccan sentence and the decoder will produce the Algerian dialect sentence. The results are given in the Table 3, unlike the results given by the Multi-task learning (One-To-Many) approach, the results in this case are mixed, we only have 50% of cases where the sequence to sequence

approach achieves better results than those given by the statistical machine translation approach. This could be explained by the lack of data necessary for learning a sequence-to-sequence neural network, while in MTL (M-2-0), the model benefited from the entire corpus for training the encoder.

	Target													
	MSA		**ALG**		**ANB**		**TUN**		**PAL**		**SYR**		**MAR**	
Source	S-Task	SMT	S-Task	SMT	S-Task	SMT	S-Task	SMT	S-Task	SMT	S-Task	SMT	S-Task	SMT
MSA	———	———	14.45	13.55	12.22	12.54	14.55	20.03	21.96	42.46	18.94	21.38	15.56	20.02
ALG	15.83	15.51	———	———	20.89	61.31	11.23	09.67	16.06	10.61	10.92	7.29	12.94	10.22
ANB	13.97	14.44	22.08	76.31	———	———	12.42	09.08	11.63	10.12	15.27	07.52	14.71	10.00
TUN	18.65	25.99	14.42	09.89	12.32	09.34	———	———	18.24	22.55	17.83	13.52	13.76	14.37
PAL	20.13	40.48	13.31	11.28	13.45	09.53	15.49	17.93	———	———	20.14	23.29	14.98	16.08
SYR	19.17	24.14	14.23	07.57	14.66	07.50	12.43	13.67	17.36	26.60	———	———	14.22	09.93
MAR	15.32	24.93	15.43	10.13	15.76	10.16	14.77	14.68	18.19	18.91	17.01	09.68	———	———

Table 3. Comparison of the performance of the S-task learning model and the statistical model of [16]

Performance of the One-To-Many multi-task learning approach In this experiment, we test an end-to-end architecture with one encoder, and one decoder for each target language. It is worth pointing out that the authors of [16] trained one statistical model, and then translated all the other dialects with the learnt translation model. From this point of view, the One-To-Many (O-2-M) multitask learning approach is thus somehow similar to the method in [16]. In Table 4, we report the results of the 0-2-M architecture and we compare them to those given by the statistical approach. We notice that the One-To-Many model is more efficient than the statistical model in 62% of cases. We can in particular notice that the Algerian dialect is better translated by the O-2-M approach than by the statistical approach with the exception of the couple of dialects Algiers-Annaba, two dialects of the same country and sharing more than 60% of words.

	Target													
	MSA		**ALG**		**ANB**		**TUN**		**PAL**		**SYR**		**MAR**	
Source	O-2-M	SMT	O-2-M	SMT	O-2-M	SMT	O-2-M	SMT	O-2-M	SMT	O-2-M	SMT	O-2-M	SMT
MSA	———	———	16.07	13.55	16.31	12.54	15.59	20.03	23.66	42.46	17.60	21.38	17.19	20.02
ALG	17.05	13.55	———	———	21.70	61.06	13.87	09.67	16.75	10.61	15.42	07.29	16.30	10.22
ANB	14.96	14.44	21.82	67.31	———	———	14.06	09.08	16.94	10.12	14.12	07.52	14.27	10.00
TUN	17.64	25.99	13.66	09.89	14.75	09.34	———	———	18.69	22.55	15.84	13.05	15.95	14.37
PAL	21.60	40.48	15.16	11.28	14.75	09.53	16.85	17.93	———	———	18.79	23.29	16.54	16.08
SYR	18.20	24.14	13.84	07.57	14.50	07.50	14.70	13.67	19.71	26.60	———	———	15.29	09.93
MAR	18.21	24.93	16.48	10.13	13.85	10.16	13.75	14.68	18.02	18.91	15.61	09.68	———	———

Table 4. Comparison of the performance of the O-2-M Multi-task learning model and the statistical model of [16]

In order to go beyond the BLEU score, we give in Figure 5 an example of translations from Algiers dialect to Annaba dialect using single task, One-to-Many and Many-to-One models. We can remark that the quality of the proposed MTL (M-2-0) translation approach is much better than the other models and this the case of all the majority of the examples we examined. Note that, in order to understand the the weakness and the mistakes carried out by the different approaches for non-Arabic speakers, we provide a word-by-word translation to English language for the source and the model's outputs.

Source (ALG)	انا البارح سمعت عبدو يهدر قال لي يديكوديو ليميساج لي يحيو للوالي حوايج كيما هاكا
Reference(ANB)	انا لبارح تسمع في عبدو يهدر قال لي يديكوديو ليميساج لي يحيو للوالي حوايج كيما اك
English translation	I yesterday heard Abdou talking he said those who are decoding the messages received by the governor something like that
Translation(S-task)	انا البار ح را ح يجي خلاه قال لي في اديكا لي لي دايما لي كيما
Translation	I yesterday will be a disaster, he said that that as usual
Translation(One-to-Many)	انا البارح يجي خلاه عبدو قال لي في اديكا لي لي كيما هاكا
Translation	I yesterday will be a disaster Abdou said in that that like
Translation(Many-to-One)	انا لبارح نسمع عبدو يهدر قال ليميساج لي يحيو للوالي حوايج كيما
Translation	I yesterday am hearing Abdou talking he said the messages received by governor things like

Table 5. An example of translations produced by different model architecture alongside the ground truth translation

6 CONCLUSION

In this article, we used a neural machine translation system for six Arabic dialects using a multitasking learning approach to tackle the problem of parallel data scarcity. The problem for Arabic dialects is the unavailability of parallel corpora for these vernacular languages. All the models we have presented are trained on a small corpus (PADIC) which is composed of only 6400 parallel sentences. The multitasking learning approach makes it possible by taking advantage of the similarity between Arabic dialects. One-to-many and many-to-one multitasking learning strategies were investigated and compared to single-task learning and statistical machine translation methods. Single-task learning and statistical methods achieved globally similar results, while One-to-many performs better in 61% of the cases in comparison to statistical approach and Many-to-one provides good quality translations and performs better in 88% of the cases. We showed,

in this article that even with a small parallel corpus, it is possible to develop neural machine translation for difficult ”languages” like Arabic dialects.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Baniata, L.H., Park, S., Park, S.B.: A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. Applied Sciences **8**(12), 2502 (2018)
3. Baniata, L.H., Park, S., Park, S.B.: A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). Computational intelligence and neuroscience **2018** (2018)
4. Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M., Shoul, M.: Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In: International Arab conference on information technology. pp. 1–6. Elsevier Science Inc New York, NY (2010)
5. Chiang, D., Diab, M., Habash, N., Rambow, O., Shareef, S.: Parsing arabic dialects. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
6. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
7. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167 (2008)
8. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: An overview. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 8599–8603. IEEE (2013)
9. Durrani, N., Koehn, P.: Improving machine translation via triangulation and transliteration. In: Proceedings of the 17th Annual conference of the European Association for Machine Translation. pp. 71–78 (2014)
10. Erdmann, A., Habash, N., Taji, D., Bouamor, H.: Low resourced machine translation via morpho-syntactic modeling: the case of dialectal arabic. arXiv preprint arXiv:1712.06273 (2017)
11. Ferguson, C.A.: Diglossia. word **15**(2), 325–340 (1959)
12. Hamdi, A., Boujelbane, R., Habash, N., Nasr, A.: The effects of factorizing root and pattern mapping in bidirectional tunisian-standard arabic machine translation (2013)
13. Harrat, S., Meftouh, K., Abbas, M., Smaïli, K.: Grapheme To Phoneme Conversion - An Arabic Dialect Case. In: Spoken Language Technologies for Under-resourced Languages. Saint Petesbourg, Russia (May 2014)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Long, M., Wang, J.: Learning multiple tasks with deep relationship networks. arXiv preprint arXiv:1506.02117 **2**, 1 (2015)
16. Meftouh, K., Harrat, S., Smaïli, K.: PADIC: extension and new experiments. In: 7th International Conference on Advanced Technologies ICAT. Antalya, Turkey (Apr 2018)
17. Meftouh, K., Bouchemal, N., Smaïli, K.: A study of a non-resourced language: the case of one of the algerian dialects. In: The third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU'12. pp. 1–7 (2012)
18. Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Smaili, K.: Machine translation experiments on padic: A parallel arabic dialect corpus. In: The 29th Pacific Asia conference on language, information and computation (2015)

19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)
20. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
21. Tachicart, R., Bouzoubaa, K.: A hybrid approach to translate moroccan arabic dialect. In: 2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14). pp. 1–5. IEEE (2014)