Machine Translation for Zero and Low-resourced Dialects using a New Extended Version of the Dialectal Parallel Corpus (Padic v2.0)

Mohamed Lichouri
Algiers, Algeria
medlichouri@gmail.com

Mourad Abbas
High Council of Arabic Language
Algiers, Algeria
abb.mourad@gmail.com

Abstract

In this paper we present a set of experiments performing machine translation related to low-resourced Arabic dialects in addition to a zero-resourced dialect (Berber). For this, we extended the parallel PADIC corpus by adding the Berber dialect corpus and translating manually more than 6000 Arabic sentences. We applied both Rule-based Machine Translation (RBMT) and Statistical Machine Translation (SMT) with and without a transliteration process. The average overall BLEU score is 42.68% with RBMT and 61.94% with SMT.

1 Introduction

Over the past years, research has seen remarkable progress on dialectal processing of the Arab region (Darwish et al., 2021), like dialect identification in text (Abbas et al., 2019; Lichouri and Abbas, 2020; Lichouri et al., 2021) and speech (Ali et al., 2021). This can be considered as a big challenge since all Arabic dialects have been spoken in the past and rarely written unlike Modern Standard Arabic (MSA). This makes processing very difficult in the absence of needed resources. This challenge is multiplied when it comes to deal with certain vernacular dialects (Lichouri et al., 2018), because they are not considered as Arabic dialects due to the obvious difference with Arabic on one side, and that they have never been written on the other side, which is the case of Berber dialects.

In this paper, we introduce PADIC v2.0, a recent version that we extended from PADIC v1.0 [1] (Meftouh et al., 2015), enriching it with new parallel texts related to a zero-resourced Berber dialect: Kabyle. To our knowledge, this is the first time that resources are developed for such a vernacular, zero-resourced dialect, and devoted to NLP and particularly machine translation. The first study that seems to be necessary and obvious to do is calculating the closeness between Berber dialect and the other Arabic dialects, Maghrebi and Levantine ones. As a natural extension to the previous studies that used PADIC (Harrat et al., 2014, 2015; Meftouh et al., 2015), we focus in this paper mainly on experiments of machine translation between Berber (Kabyle) Dialect and Arabic (MSA), as well as between the remaining Arabic dialects. The rest of this article is organized as follows, we first present related work in section 2. In section 3, we describe how we enriched PADIC corpus, followed by measuring distances between the different dialects. In section 4, we present the evaluation methods and the experimental results, and finally, we conclude in section 5.

2 Related Work

Low-resource and zero-resource languages are considerably lacking in works especially on Machine Translation (MT). For instance, for Arabic Language and its dialects, most of the work done on MT, focused on translation into English, as in (Sawaf, 2010) where an hybrid approach between rule-based and statistical methods was presented. The authors evaluated their approach on the NIST MT08 WB Arabic dataset comprising MSA and 15 colloquial Arabic dialects from almost all the Arab countries (except Algeria). In (Salloum and Habash, 2011), the authors proposed a technique to solve the problem posed by out-of-vocabulary (OOV) words and low frequency words in Arabic-English SMT. For that they adopted a paraphrasing approach of (OOV) words in Arabi Dialectal Text to produce Modern Standard Arabic (MSA) paraphrases of dialectal Arabic that are input to a phrase-based SMT system. This approach permitted the authors to implement Elissa which is

[1] PADIC v1.0 is a parallel Arabic multi-dialectal textual corpus composed of six Arabic dialects: Syrian, Tunisian, Moroccan, Palestinian and two Algerian dialects (of Algiers and Annaba cities), in addition to MSA.
https://sites.google.com/site/torjmanepnr/Home
https://sourceforge.net/projects/padic/

Arabic dialect into English Translator by pivoting on MSA (Salloum and Habash, 2012, 2013). (Zbib et al., 2012) conducted MT for (MSA-English), (Levantine-English), and (Egyptian-English). The authors found surprisingly that translating from Egyptian and Levantine dialects into English outperformed the couple of language (MSA-English) by 6.3 and 7.0 of BLEU, respectively. (Sajjad et al., 2013) attempted to narrow down the gap between Egyptian and MSA by applying an automatic character-level transformational model that changes Egyptian to a format similar to MSA, which reduced the out-of-vocabulary (OOV) words from 5.2% to 2.6% and gives a gain of 1.87 BLEU points. For Iraqi Dialect, in order to resolve the lack of dialectal parallel data, authors presented in (Kirchhoff et al., 2015) how they extracted parallel data from out-of-domain corpora related to different Arabic dialects and MSA. By applying deep neural network on Machine Translation, (Zoph et al., 2016) presented an approach based on transfer learning for the benefit of low-resource languages. In another context, (Almahairi et al., 2016) conducted experiments on Arabic Neural Machine Translation (NMT) and have concluded that in spite of the big need of tremendous amount of data for NMT, in comparison to Phrase-based Statistical Machine Translation, the NMT system outperforms the statistical one in case of an out-of-domain test set, making it attractive for real-world deployment. A comparison between statistical and NMT was conducted in (Guellil et al., 2017) for MSA and one of its dialects that had been extracted from PADIC corpus. Another study presented in (Alrajeh, 2018), having the same objective as that mentioned in (Guellil et al., 2017), which is comparing between phrase-based SMT and NMT, has been reached using three parallel MSA-ENG corpora: UN, ISI and Ummah. Their findings show that tuning a model trained on the whole data using a small high quality corpus like Ummah gives a substantial improvement and that training a neural system with a small Arabic-English corpus is competitive to a traditional phrase-based system. Another aspect that is not taken into account by most current models is that a sentence can have multiple translations. For this, as to solve the problem of this kind of variation in parallel corpus, (Schulz et al., 2018) applied a deep generative model of machine translation which incorporates a chain of latent variables, in order to account for local lexical and syntactic variation in parallel corpora.

3 Description of PADIC v2.0

The difference between PADIC v1.0 and PADIC v2.0 is that PADIC v2.0 has been enriched with Kabyle (an Algerian zero-resourced dialect). Kabyle is one of the Berber dialects; it is a branch of the Afro-Asiatic language phylum which covers parts of North Africa, stretching from Morocco to Yemen, including Libya, Egypt and Somalia. In the following, we will present how we developed PADIC v2.0, as well as the linguistic similarities between Arabic and Berber that can be on all levels: phonology, morphology, syntax, and lexicon.

3.1 Enrichment of PADIC with Berber Dialect

We solicited a couple of native speakers of Kabyle, a variant of Berber dialect, from Tizi-Ouzou city. These native speakers translated the 6400 sentences of PADIC from Algiers dialect (ALG), writing these sentences with Arabic letters. Hence, the new PADIC version is composed of seven Arabic dialects: ALG (Algiers), ANB (Annaba), TUN (Tunisia), PAL (Palestine), SYR (Syria), MOR (Moroccan), and KAB (Kabyle), in addition to MSA.

3.2 Distances between Dialects

In order to quantify the closeness between the studied dialects, we used a set of distances belonging to five measure classes[2]:

Edit based: Hamming, MLIPNS, Levenshtein, Damerau-Levenshtein, Jaro-Winkler, Strcmp95, Needleman-Wunsch, Gotoh and Smith-Waterman (Navarro, 2001).

Token based: Jaccard index, Overlap coefficient and Cosine similarity.

Sequence based: Longest common substring similarity and Ratcliff-Obershelp similarity.

Phonetic: MRA and Editex.

Compression NCD-based: BZ2, LZMA and ZLib.

The choice of these measures is explained by the fact that each of them has its own calculation algorithms and therefore each has a specific

[2]https://pypi.org/project/textdistance/

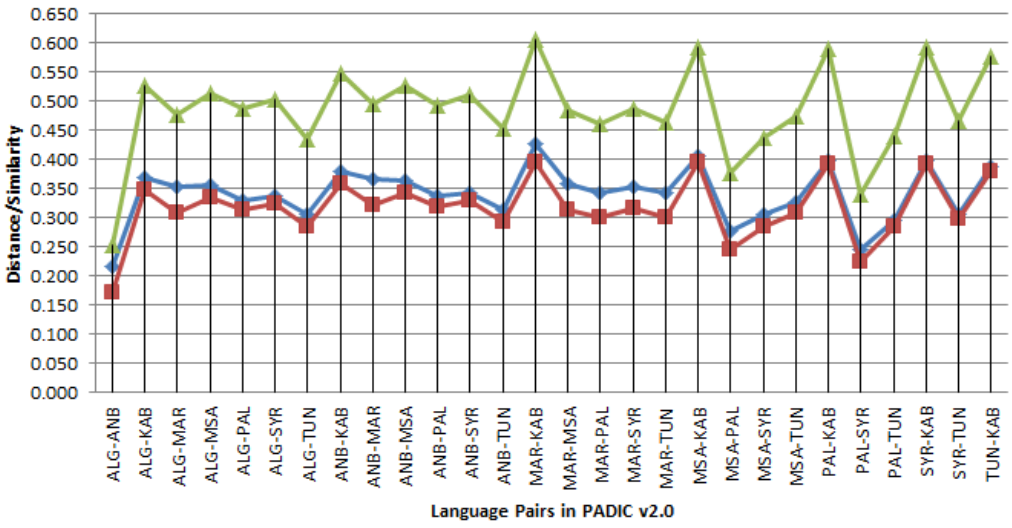

Figure 1: Distance Measures between Language Pairs in PADIC v2.0 by Compression NCD-based metric(0=Equal,1=Different). Metrics Bz2-NCD (Blue), LzMA-NCD (Red) and zLIB-NCD (Green).

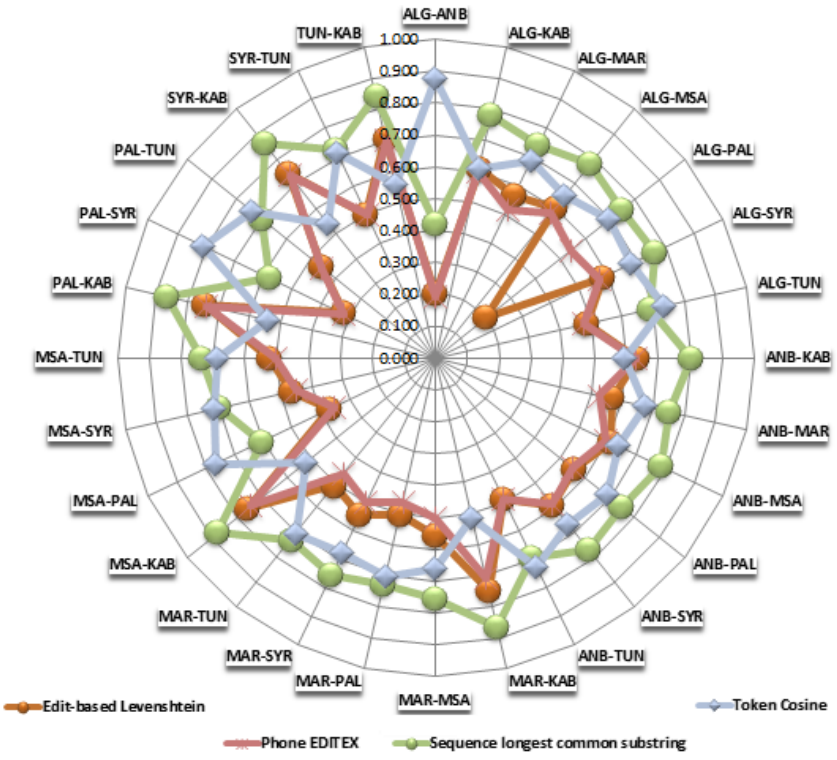

Figure 2: A sample of Distance Measures between Language Pairs in PADIC v2.0 by Edit-based, Phonetic, Sequence-based and Token-based metrics (0=Equal,1=Different)

purpose to deploy[3]. Because of the variation of these different measures, we used the normalized similarity for sequences (Vitányi, 2011), that returns a float between **0** and **1** (Cilibrasi and Vitányi, 2005), where **1** means totally different, and **0** means equal[4]. Based on Figure 1, we can see that the results obtained using Bz2-NCD and LzMA-NCD are very close, and 0.15 lower than zLIB-NCD. Note that the three curves have relatively a similar behavior, this reinforces the differences and similarities that actually exist between these dialects.

[3]https://www.kdnuggets.com/2019/01/comparison-text-distance-metrics.html

[4]https://articles.orsinium.dev/p/notes-other/ncd/

	ALG	ANB	KAB	MOR	PAL	SYR	TUN
From MSA (Simple)	65.93	65.93	62.91	71.24	73.72	71.76	71.67
From MSA (Translit)	09.56	09.63	09.44	09.63	09.67	09.41	09.63
To MSA (Simple)	69.56	70.29	64.06	80.14	72.12	64.63	57.56
To MSA (Translit)	15.01	16.87	33.81	14.56	39.32	16.65	30.40

Table 1: Comparison results of Rule-based MT for PADIC v2.0 from/to MSA with and without transliteration process.

	ALG	ANB	KAB	MOR	PAL	SYR	TUN
From MSA (Simple)	52.51	47.55	35.52	60.36	81.12	70.39	59.91
From MSA (Translit)	57.66	54.63	44.89	65.52	87.35	73.71	68.81
To MSA (Simple)	71.71	70.46	25.53	66.53	83.05	69.45	72.16
To MSA (Translit)	51.62	51.37	25.89	63.84	88.96	69.97	63.96

Table 2: Comparison results of SMT for PADIC v2.0 from/to MSA with and without transliteration process.

For example, Algiers dialect (ALG) is the closest to the Algerian Annaba's dialect (ANB), Which is very reasonable and expected, since these two dialects are spoken in the same country and share up to 60% of words (Meftouh et al., 2015). However, some results are unforeseen if one takes into account the geographical parameter. The appropriate example for this case, is the pairs of Arabic dialects TUN/ALG and TUN/PAL. Indeed, Tunisian has small distances with both Algiers dialect and Palestinian. The closeness with ALG is understandable because Tunisia borders Algeria, which is not the case for Palestine. Another interesting and unexpected result is that Moroccan dialect (MOR) is closer to Palestinian than Tunisian or Algerian, though Morocco borders Algeria. For Levantine dialects, Syrian is close to Palestinian, which is not surprising because of the geographical proximity, whereas Palestinian is closer to MSA than Syrian. For the Berber variant (Kabyle), it is clearly shown in Figure 1, that it has the farthest distance with all the Arabic dialects.

4 Experiments and Results

We applied two well-known MT approaches using PADIC v2.0: Rule Based MT (RBMT) and Statistical MT (SMT). We decided to use two versions of PADIC, one with Arabic letters, and the other one by applying Buckwalter transliteration [5]. The results are evaluated using BLEU score [6].

4.1 Rule Based MT

For achieving a simple rule based machine translation, we adopted the same model used in in the work by Niyongabo & College [7]. The obtained results are presented in table 1. The best BLEU scores are obtained for the couples (MSA-Pal) (73.72%) and (MOR-MSA) (80.14%). In general, translation from MSA into (MOR, PAL, SYR, ANB) yielded close BLEU scores, around 71%, and from MSA into Algerian dialects (Alg, ANB, Kab) the scores are around 63%. Whereas, we recorded surprisingly, the same BLEU score (around 64%) when translating into MSA,from KAB and SYR. On the other hand, (ALG, ANB, PAL) have an overall score of 69%. The worst BLEU for Translation into MSA is the one recorded from ANB: 57.56%. The impact of transliteration was very negative, the BLEU score is around 9.5% for (dialects-MSA) and ranges for (MSA-dialects) from 15% to 23.8% .

4.2 SMT

We used Moses2.0 to train the SMT model. The obtained results are presented in table 2. We can say that without transliteration, the performance of translation from MSA achieved by SMT is lower than RBMT, except for Palestinian (MSA-PAL) that has the best BLEU score (81.12%). However, Translation into MSA using SMT outperforms RBMT except for Kabyle dialect (MSA-KAB) Contrarywise, as shown in Table 2, the Buckwalter transliteration has a positive impact on SMT performance in most of cases except for the three

[5] http://www.qamus.org/transliteration.htm

[6] https://github.com/cshanbo/Smooth_BLEU/blob/master/BLEU.py

[7] https://github.com/pniyongabo/kinyarwandaRBMT

pairs (ANB-MSA), (ALG-MSA) and (ANB-MSA) . Note that the best BLEU score obtained in our experiments is 88.96% for (PAL-MSA). On the other hand, SMT provides the worst results for Kabyle dialect (44.89% for MSA-KAB) and (25.89% for KAB-MSA), though the rule based method yielded promising and surprising results (62.91% for MSA-KAB) and (64.06% for KAB-MSA).

5 Conclusion

In this paper, we presented a new extension to the Parallel Arabic Dialect Corpus PADIC by adding a zero-resourced dialect, namely: Algerian Kabyle dialect. We tested rule based and statistical machine translation models. We studied the impact of using Buckwalter transliteration on the performance of the trained models. The results are promising, we believe that we can further enhance the performance by applying some preprocessing steps as sentence tokenization, and using Neural MT.

References

Mourad Abbas, Mohamed Lichouri, and Abed Alhakim Freihat. 2019. St madar 2019 shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 269–273.

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting arabs: bridging the gap in dialectal speech recognition. *Communications of the ACM*, 64(4):124–129.

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron C. Courville. 2016. First result on arabic neural machine translation. *CoRR*, abs/1606.02680.

Abdullah Alrajeh. 2018. A recipe for arabic-english neural machine translation. *CoRR*, abs/1808.06116.

Rudi Cilibrasi and Paul MB Vitányi. 2005. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Imane Guellil, Faical Azouaou, and Mourad Abbas. 2017. Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter.

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, and Kamel Smaili. 2015. Cross-dialectal arabic processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 620–632. Springer.

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

Katrin Kirchhoff, Bing Zhao, and Wen Wang. 2015. Exploiting out-of-domain data sources for dialectal arabic statistical machine translation. *arXiv preprint arXiv:1509.01938*.

Mohamed Lichouri and Mourad Abbas. 2020. Simple vs oversampling-based classification methods for fine grained arabic dialect identification in twitter. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 250–256.

Mohamed Lichouri, Mourad Abbas, Abed Alhakim Freihat, and Dhiya El Hak Megtouf. 2018. Word-level vs sentence-level language identification: Application to algerian and arabic dialects. *Procedia Computer Science*, 142:246–253.

Mohamed Lichouri, Mourad Abbas, Khaled Lounnas, Besma Benaziz, and Aicha Zitouni. 2021. Arabic dialect identification based on a weighted concatenation of tf-idf features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 282–286.

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

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal arabic to english. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 1–6.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard arabic machine translation system. *Proceedings of COLING 2012: Demonstration Papers*, pages 385–392.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the conference of the association for machine translation in the americas (amta), denver, colorado*.

Philip Schulz, Wilker Aziz, and Trevor Cohn. 2018. A stochastic decoder for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1243–1252.

Paul M.B. Vitányi. 2011. Compression-based similarity. In *2011 First International Conference on Data Compression, Communications and Processing*, pages 111–118.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.