Neural Vs Statistical Translation of Algerian Arabic Dialect written with Arabizi and Arabic letter

Imane GUELLIL
Ecole Supérieure d'informatique
d'Alger ESI
Ecole préparatoire des sciences et techniques
d'Alger EPSTA

i_guellil@esi.dz

Faical AZOUAOU
Ecole Supérieure d'informatique
d'Alger ESI

f_azouaou@esi.dz

Mourad ABBAS
Centre de Recherche Scientifique et Technique pour le
Développement de la Langue Arabe (CRSTDLA)
m_abbas04@yahoo.fr

Abstract

The arabizi is the combination between the latine letters and the number in the same word. The transliteration is the process of transforming the text from one alphabet to another. In our case, it consists to transform arabizi to Arabic. The transliteration is considered as the first compound of automatic translation when users combine between different alphabets to write.

Almost all the work on Arabic Dialect Translation is based on statistical translation and we did not get any work on Neural translation. To bridge this gap, we propose in this paper an approach composed of two steps: 1) Transliteration and 2) Translation. In the two steps we call the neural network and compare the results with traditional statistical approach. We apply our approach on dialect spoken by over 40 million people and suffer from the lack of works that handle it. This dialect is Algerian Dialect and it is recognized by a high level of combination between Arabic and arabizi.

Experimental results show that neural approach gives better results for transliteration when statistical approach stay gives better results for translation. However, these results could be justified by the fact of the minimal number of our parallel corpus (only 6412 sentences)

1 Introduction

Machine Translation (MT) represents an active research area (Chand, 2016). The most used approaches for MT are respectively rule-based and Statistical Machine translation (SMT) (Antony, 2013; Alqudsi et al., 2014). Recently, a new approach has emerged which involves neural networks. This approach is known as Neural Machine Translation (NMT) (Sutskever et al., 2014; Cho et al., 2014a; Bahdanau et al., 2014). NMT has not been applied on Arabic and its dialects yet except in (Almahairi et al., 2016) where only Modern Standard Arabic (MSA) has been dealt with.

We note that almost all works on Arabic dialects are based on messages provided from social media. In fact, social media users write Arabic dialects in two different ways:

1) By using only Arabic letters: for example, " حبيت من فضلكم6نسقسيكم شحال يدير ايفون", which means, "I want to ask you, what is the price of iphone6 please, or " عفسة مليحة", which means "a good thing". These sentences belong to a corpus

containing 18603 sentences collected by ourselves from Facebook on January 2014.

2) By combining Latin letters and numbers: for example: “walahi rabi ykon fi el3awn” which means: “I swear god will help you”. These messages were extracted from an Algerian Arabic-French code switched corpus of Cotterell (Cotterell et al., 2014).

This way of writing takes different names like “Franco-Arabic”, “Romanized Arabic”, “Arabizi”, “Arabish”,… etc, (Chalabi and Gerges, 2012). In this paper, we use the term “Arabizi”, based on the work of Darwish (Darwish, 2014), which proposes an approach to identify and transliterate Arabizi. The work in (Bies et al., 2014) considers Arabizi as a challenge for Arabic NLP research. To address this challenge, we consider Arabizi Transliteration as the first module (or as pre-processing step) of Arabic dialect treatment of the analyzed messages combining between the Arabizi and the Arabic letters. We survey a lot of work on Statistical Machine transliteration (SMTR) (Van der Wees et al., 2016; Al-Badrashiny et al., 2014; Darwish, 2014) and other combining transliteration and translation (May et al., 2014 ; Van der Wees et al., 2016). However, we should note that no work related to Neural Machine Transliteration (NMTR) of Arabizi has been achieved. To address this problem, we propose an Arabic dialect translation system composed of two components: The first one for Arabizi transliteration and the second one for Arabic dialect translation. The Arabizi transliteration is based on the two models: Statistical and Neural. Arabic translation could be done on Arabizi or on Arabic dialect written with Arabic letters. This component is also based on both statistical and neural models.

The system is focused on an under-resourced language variant, namely Algerian dialect, which belongs to the Maghrebi Dialect family. Only few pieces of work have been done on the Algerian Dialect (Cotterell et al., 2014; Meftouh et al., 2015). Concerning the data set, we used PADIC, a multi Parallel Arabic DIalect Corpus (Meftouh et al., 2015).

The present paper is organized as follows. In section 2, we review previous work based on translation and transliteration of Arabizi, we also highlight the work combining the two tasks. In section 3, we describe our proposed approach. In section 4, we present our experiments and results. We then conclude in section 5 where we also present a set of perspectives.

2 Related work

By following the survey of (Shoufan and Ameri 2015), we can regroup works on Arabic dialect into four families: 1) Basic analysis, 2) Resources construction, 3) Identification and 4) semantic analysis. Lot of works have been done in each category. Basic analysis consists of orthographic, syntactic and Part of speech analysis (Guellil and Azouaou 2017a; Habash and Rambow 2006). Resources construction consists of the construction of lexicon and corpora (Meftouh et al. 2015). Identification is the process of determination the type of dialect in a set of dialect (Guellil and Azouaou 2016). Semantic analysis means Automatic translation (MT) and sentiment analysis (Guellil and Boukhalfa 2015).

In this section, we survey the work that has been done on MT by distinguishing between the work on Arabic language and dialects and those done on other languages. We also present the work on transliteration, by considering Arabizi transliteration. We finish by discussing works that deal with transliteration for the purpose of translation, which is our goal.

2.1 The works on Machine Translation

Based on the two surveys of (Antony, 2013) and (Alqudsi et al., 2014), the most famous approaches of MT are the rule-based (Salem et al., 2008) and the statistical approach (Lopez, 2008; Och and Ney, 2004). Moreover, statistical phrase-based MT is the most used statistical approach, because it maximizes the translation probability (Koehn et al., 2003 ; Costa-Jussa and Fonollosa, 2016). Some researches combine the two approaches (rule-based and statistical) in a hybrid translation model (Langlais and Simard, 2002; Groves and Way, 2005). Some of these works concern Arabic (Alqudsi et al., 2014 ; Salem et al., 2008).

A lot of works have been done also on NMT (Sutskever et al., 2014; Cho et al., 2014a; Bahdanau et al., 2014). For Arabic, it has been presented in (Almahairi et al., 2016), the first results concerning NMT in the two directions

(Arabic to English and English to Arabic). For Arabic dialects translation, many works have been carried out (Sadat et al., 2014; Zbib et al., 2012; Sawaf, 2010). However, there is no work that investigated NMT for these under-resourced dialects.

2.2 The work on Machine Transliteration

There are many works on transliteration that have been achieved for languages like: Hindi, Punjabi or Chinese. Other works have been done on Arabic (Habash et al., 2007; Al-Onaizan and Knight, 2002).

Almost all works handling Arabizi transliteration considered this problem as a Statistical Machine Translation problem, based on the character level (Van der Wees et al., 2016; Al-Badrashiny et al., 2014; Darwish, 2014). In these works, all sentences have been divided to a set of words, so a word is considered as a sentence. Then, each word has been divided to a set of characters. Each character is considered as a word.

Recently, transliteration has taken a new direction, where neural networks have been applied (Shao and Nivre, 2016; Jadidinejad, 2016; Rosca and Breuel, 2016; Guellil et al 2017b). In (Shao and Nivre, 2016), a comparison between NMTR and SMTR for Chinese-English transliteration. In (Jadidinejad, 2016), the authors present a character-based model for NMTR. They obtained better results than baseline, which is calculated using "Moses" toolkit (Koehn et al., 2007). However, the baseline does not contain Arabic or Arabizi. In (Rosca and Breuel, 2016), sequence to sequence models have been used. Transliteration has been achieved on three pairs of languages: English-Japanese, English-International Phonetic Alphabet (IPA) and English-Arabic.

2.3 The work combining between transliteration and translation

In (May et al., 2014) and (Van der Wees et al., 2016), a Statistical Machine Translation from Arabizi to English has been presented. In the first work a transliterated corpus has been constructed semi-automatically, and weights of characters are learned from an Arabizi-Arabic text. While in the second work the corpus has been built automatically, and uniform weights have been used. However, there is no work in which NMTR and SMT or NMT have been combined for Arabizi.

Our contribution in this paper is firstly to present results concerning NMTR from Arabizi to MSA and compare them to results obtained by SMTR. We also present the SMT and the NMT results of the transliterated messages.

3 The Arabic dialect translation framework

The main purpose of this paper is to present an approach for Arabic dialect translation. We highlighted the fact that social media users write in different manners (with Arabic letters and Arabizi). Before making this translation, we have to transliterate an Arabizi text (or comments, in the case of social media) to Arabic script. As a consequence, the general idea of this approach is to transliterate an Arabizi corpus with SMTR and NMTR techniques and translate the transliterated texts into MSA.

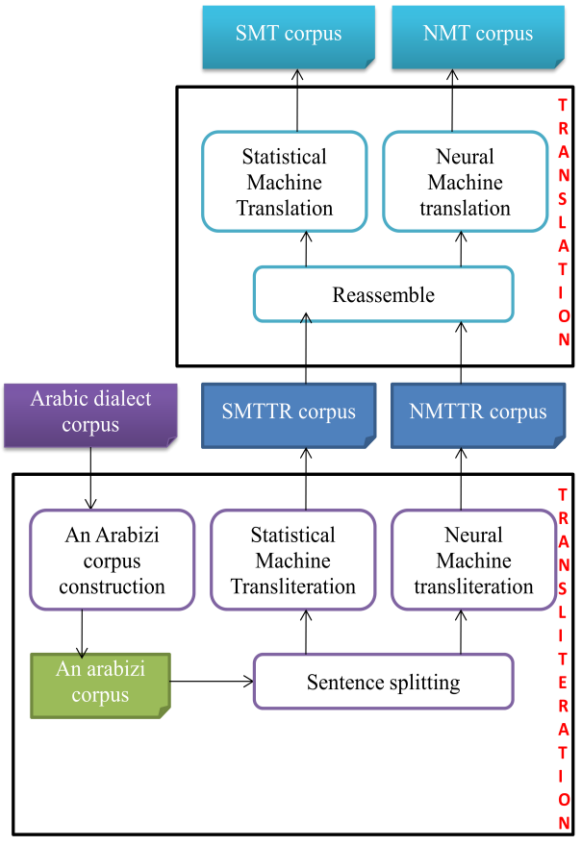

Fig.1: Arabic dialect translation framework

3.1 The transliteration step

The main idea of this step is to transliterate a given text written in Arabizi to the same text written in the Arabic script. To achieve this task, we follow four main sub-steps:

1) We construct a parallel Arabizi corpus containing 6233 sentences. We used PADIC (which is written in Arabic letters), and we transliterated it to Arabizi. We achieved the transliteration by defining a rule-based algorithm to automatically transliterate Arabic Dialect written in Arabic letters to Arabizi form. This algorithm transforms the letter (ع) to the number (3), the letter (غ) to the two letters (gh),…etc. For establishing this convention or rules, we analyze a set of Arabic dialect messages. We finish by manually reviewing the transliteration. At this stage, we could only correct 1300 sentences.

2) Based on the work of (Darwish, 2014), we divided each sentence to a set of word and each word to a set of characters, so we work at the character level.

3) We apply a phrase based SMT on our data. These data are first trained using a language model. The language model is built with the target language (in our case, Arabic Dialect written with Arabic letters). For training the transliteration model, we run a character based alignment. We finish by the tuning process, for determining the best results for each transliteration pair.

4) We also Apply to the same data an NMT model. In this paper, we used RNN Encoder-decoder model proposed by (Cho et al., 2014a) and (Sutskever et al., 2014). The choice of RNN Encoder-decoder is mainly due to the fact that it is considered as the simplest version of neural machine translation. The idea of the Encoder-decoder is to read word by word from the input sentence (in our case character by character of the input word), and encode the words to a sequence of hidden states. Then the decoder computes all the possible translations based on the context (in our case all transliterations) and generates the corresponding translation (transliteration) (Jadidinejad, 2016; Kikuchi et al., 2016; Cho et al., 2014b). To train this model, we firstly replace some unknown characters by the term “unk”. We used a development set separated from the training set to measure how well the model generalizes during training. Finally we use an external lexicon indicating the mapping between characters and their probabilities. To create this lexicon, we used a word alignment tool (character-based) (Neubig, 2016).

3.2 The translation step

The main idea of this step is to translate Arabic Dialect into MSA. This allows us, in the future, to consider MSA as a pivot for translating into English or French. We assume that each sentence is written in Arabic letters only or Arabizi only. We do not treat, in this paper, the case where we find an Arabic letter and Arabizi in the same sentence. We leave this problem for future research. This component could take as an input arabizi messages after transliteration or the messages written with Arabic letters. So it can receive as the input messages provided from our Arabic dialect corpus or a set of messages that we transliterated before (the output of the transliteration component). In this step, we follow three main sub-steps.

1) We begin by reassembling the words of the transliterated corpus. This is due to the fact that transliteration is word-based level and translation is phrase-based level. However, we do not need to reassemble in the case of Arabic dialect translation (when corpus is written with Arabic letters), as shown in Figure 1.

2) We apply an SMT model to the resulting sentences. As in the transliteration task, we have to build the language model, train it by running a word-level alignment and call the tuning process.

3) We apply an NMT model to the same sentences. We also use the RNN Encoder-decoder model. We follow the same steps as the transliteration, so we detect the unknown words and train the model and create an aligned lexicon. The only difference compared to transliteration is that the model is phrase-based and not word-based.

4 Experiments and results

Our System is composed of two components: transliteration and translation, for which we applied statistical and neural models. Concerning the statistical model, we used Moses toolkit, with KenLM (Heafield, 2011) for language modeling

and GIZA++ for alignment (Och and Ney, 2000). Concerning Neural model, we used Lamtram toolkit (Neubig, 2015), which is the combination of the of the two aforementioned models (Bahdanau et al., 2014; Luong et al., 2015). Before using lamtram toolkit, we had to install DyNet library, formerly known as CNN, developed by Carnegie Mellon University.

We tested our Approach on Algerian Dialect, so we the Algerian and MSA parts of PADIC. After having built our transliterated corpus, we manually checked it, so we could review 1300 sentences. Based on these sentences, we constructed different training, development and test corpora to test the transliteration algorithm.

After testing the transliteration, we applied SMT and NMT techniques described below (on the same test corpora). We present in the Table 1 the different results concerning the different combination of SMTR/ NMTR and SMT/NMT

Training size Sentences/ words	**Accuracy level**	**SMTR**	**NMTR**
100/1078	Character	71.47	67.24
	Word	57.81	59.63
250/2208	Character	76.83	75.73
	Word	65.95	69.37
500/3537	Character	78.63	78.23
	Word	67.55	70.87
1000/6444	Character	**80.01**	**80.97**
	Word	**71.48**	**73.66**

Table 1: SMTR Vs NMTR of Arabizi

We clearly observe that Neural transliteration gives better results than Statistical transliteration. The best results that we achieved are on the biggest training set, 1000 sentences. Then, we kept the best transliteration (SMTR and NMTR) and translated the giving sentences to MSA. Table 2 shows the different results of NMT models evaluation.

Through Table 2, we observe that neural translation begins to give results with the maximum size of training corpus, so 1000 sentences for the transliteration step and 5500 sentences for the translation step. However, we observe that this corpus is too small, what it justifies the minimal score of translation.

We also observe that SMTR leads to improve the translation comparing the NMTR (by knowing that SMTR gave better results that NMTR, show the Table 1).

Trainig corpus size	Transliteration	Neural Translation BLEU score
10%	Reference	0.00
	SMTR	0.00
	NMTR	0.00
25%	Reference	1.71
	SMTR	0.00
	NMTR	0.0
50%	Reference	2.34
	SMTR	0.0
	NMTR	0.0
100%	Reference	6.25
	SMTR	4.54
	NMTR	4.13

Table 2: Neural Machine Translation of Arabic dialect

To compare our results, we lead the same experimentations with SMT model (used by almost all works of the state of the art). We present the obtained results on Table 3.

In contrast of NMT, SMT begins to give results with only small training corpus (10% of the general corpus size, show Table 3). We also observe that SMT didn't give the "Zero" score.

Training corpus size	Transliteration	Statistical Translation BLEU score
10%	Reference	6.31
	SMTR	2.65
	NMTR	2.40
25%	Reference	8.02
	SMTR	3.47
	NMTR	4.49
50%	Reference	10.02
	SMTR	5.21
	NMTR	4.21
100%	Reference	**10.74**
	SMTR	**6.01**
	NMTR	3.94

By comparing the Table 2 and 3, we observe that SMT gives better results than NMT. Moreover, SMT is well performing where combined with SMTR. However the work in state of the art showed that NMT gives better result that SMT for rich-resources languages like English. The best problem that we highlight for Arabic and its Dialect through these experimentations is the lack of resources. We are convinced that if we raised the training corpus size, we could improve these results.

We also study the impact of transliteration before translation, to show it, we conducted SMT on the Arabizi corpus test without making transliteration. We carried out this experiment for the biggest size of training corpus (100% of the total size). We obtained a BLEU score=4.26 where the score after SMTR=6.01 and the reference=10.74. We conclude that transliteration before translation improve the final resuts.

5 Conclusion and Perspectives

The main idea of this work is to present a system for Arabic Dialect translation. The big challenge that we face is that Arabic dialect could be written with Arabizi or Arabic letter. Before making translation, Arabizi have to be standardized. For this purpose, we presented and implemented an approach composed by two components: transliteration and translation. Through this paper, we noticed that for a small Arabizi corpus, NMTR provided better results than SMTR, whereas SMT still gives better results than NMT. This is mainly due to the fact that NMT needs a huge training set. This can be noticed in the case of small sizes of training corpus, where the BLEU score of NMT is equal 0.0.

In perspective, we aim to generalize this idea by testing our system on other corpora like that of Cotterell et al. (Cotterell et al. 2014). We also plan to merge different Arabic dialect corpora to improve statistical and neural translation.

However, our first perspective is to concentrate on resource construction and enrichment for Algerian Dialect.

Acknowledgments

We would like to thank different researcher for having send us their lexicons and corpora. Between them: Rayan Cottrel for its code switched corpus, Gilbert Badaro for ArSenl lexicon, Pierre charron for NRC corpus, Eshrag Refaee for its Twitter annotated corpus and Khaled Elmiman for its Multi dialect text corpora. We also would like to particularly thank Salima Harrat for its help concerning the manipulation of the tool MOSES (dedicated to statistical translation) and their different advices.

References

Al-Onaizan, Yaser, and Kevin Knight. 2002. Machine transliteration of names in Arabic text. *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, 2002, pp. 1-13. Association for Computational Linguistics.

Almahairi, A., Cho, K., Habash, N., & Courville, A. 2016. First Result on Arabic Neural Machine Translation. arXiv preprint arXiv:1606.02680.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bhalla, Deepti, Nisheeth Joshi, and Iti Mathur. 2013. Rule based transliteration scheme for English to Punjabi. arXiv preprint arXiv:1307.4300.

Chalabi, Achraf, and Hany Gerges. 2012. Romanized arabic transliteration.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cotterell, Ryan, et al. 2014. An algerian arabic-french code-switched corpus. *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*.

Darwish, Kareem. 2014. Arabizi Detection and Conversion to Arabic. *ANLP*.

Darwish, Kareem, and Walid Magdy. 2014. Arabic information retrieval. Foundations and Trends, *Information Retrieval* 7(4):239-342.

Dyer, Chris, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. *Association for Computational Linguistics.*

Guellil I, Boukhalfa K Social big data mining: A survey focused on opinion mining and sentiments analysis. In: Programming and Systems (ISPS), 2015 12th International Symposium on, 2015. IEEE, pp 1-10

GUELLIL, Imane, AZOUAOU, Faiçal, ABBAS, Mourad, et al. Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic. In : Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation (co-located with EAMT2017). 2017b.

GUELLIL, Imène et AZOUAOU, Faiçal. Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect. In : Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on. IEEE, 2016. p. 724-731.

GUELLIL, Imène et AZOUAOU, Faiçal. ASDA: Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique. arXiv preprint arXiv:1707.08998, 2017a.

Habash, Nizar, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. In *Arabic computational morphology*. pp. 15-22: Springer.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011, pp. 187-197. *Association for Computational Linguistics.*

Jadidinejad, Amir H. 2016. Neural Machine Transliteration: Preliminary Results. arXiv preprint arXiv:1609.04253.

Josan, Gurpreet Singh, and Gurpreet Singh Lehal 2010. A Punjabi to Hindi Machine Transliteration System. *Computational Linguistics and Chinese Language Processing* 15(2):77-102.

Joshi, Hardik, Apurva Bhatt, and Honey Patel. 2013. Transliterated Search using Syllabification Approach. *Forum for Information Retrieval Evaluation*, 2013.

Kaur, Kamaljeet, and Parminder Singh. 2014. Review of Machine Transliteration Techniques. International Journal of Computer Applications 107(20).

Kikuchi, Yuta, Neubig, Graham, Sasano, Ryohei, Takamura, Hiroya and Okumura, Manabu. 2016. Controlling output length in neural encoder-decoders. arXiv preprint arXiv:1609.09552.

Kingma, Diederik, and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koehn, Philipp, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177-180. Association for Computational Linguistics.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Malik, M.G. Abbas, Boitet, Christian, Besacier, Laurent and Bhattcharyya, Pushpak. 2013.Urdu Hindi machine transliteration using SMT. *WSSANLP2013*.

May, Jonathan, Yassine Benjira, and Abdessamad Echihabi. 2014. An Arabizi-English social media

statistical machine translation system. *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pp. 329-341.

Meftouh, Karima, Harrat, Salima, Jamoussi, Salma, Abbas, Mourad and Smaili, Kamel. 2015. Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus. *The 29th Pacific Asia Conference on Language, Information and Computation*.

Neubig, Graham. 2015. Lamtram: A toolkit for language and translation modeling using neural networks.

Neubig, Graham. 2016. Lexicons and minimum risk training for neural machine translation: *NAIST-CMU at WAT*2016. arXiv preprint arXiv:1610.06542.

Och, Franz Josef, and Hermann Ney. 2000. Giza++: Training of statistical translation models.

Oh, Jong-Hoon, and Key-Sun Choi. 2005. An ensemble of grapheme and phoneme for machine transliteration. *International Conference on Natural Language Processing*. 2005. pp. 450-461. Springer.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, pp. 3104-3112.

Van der Wees, Marlies, Arianna Bisazza, and Christof Monz. 2016. A Simple but Effective Approach to Improve Arabizi-to-English Statistical Machine Translation. *WNUT* 2016.

Zaidan, Omar F, and Chris Callison-Burch. 2014. Arabic dialect identification. Computational Linguistics 40(1):171-202.