Translating Dialectal Arabic as Low Resource Language using Word Embedding

Ebtesam H Almansor [1,2], **Ahmed Al-Ani** [1]
[1]Faculty of Engineering and Information Technology, University of Technology Sydney, Sydeny, Australia
[2]Community College, Najran University, Najran, Saudi Arabia
EbtesamHussain.Almansor@student.uts.edu.au
Ahmed.Al-Ani@uts.edu.au

Abstract

A number of machine translation methods have been proposed in recent years to deal with the increasingly important problem of automatic translation between texts of different languages or languages and their dialects. These methods have produced promising results when applied to some of the widely studied languages. Existing translation methods are mainly implemented using rule-based and static machine translation approaches. Rule based approaches utilize language translation rules that can either be constructed by an expert, which is quite difficult when dealing with dialects, or rely on rule construction algorithms, which require very large parallel datasets. Statistical approaches also require large parallel datasets to build the translation models. However, large parallel datasets do not exist for languages with low resources, such as the Arabic language and its dialects. In this paper we propose an algorithm that attempts to overcome this limitation, and apply it to translate the Egyptian dialect (EGY) to Modern Standard Arabic (MSA). Monolingual corpus was collected for both MSA and EGY and a relatively small parallel language pair set was built to train the models. The proposed method utilizes Word embeddings as it requires monolingual data rather than parallel corpus. Both Continuous Bag of Words and Skip-gram were used to build word vectors. The proposed method was validated on four different datasets using a four-fold cross validation approach.

1 Introduction

Globally, social media networking platforms have witnessed a rapid increase in the last few years (Albogamy and Ramsay, 2015). Social media messages usually contain large amounts of noisy text. Thus, issues of the noisy text generation are increasing. A noisy text is an informal text that contains spelling error, slang, dialects and abbreviation (Li and Liu, 2012). Volumes of informal texts require efficient processing and analysis techniques such as sentiment analysis and summarization (Han and Baldwin, 2011). Also, these noisy texts need to be translated to their standard form to be more understandable. Therefore, various studies in Natural Language Processing (NLP) were focused on translating these texts (Han and Baldwin, 2011). This work explores dialectal Arabic translation.

Dialect Arabic words can be treated as non-standard words that are used in Arabic and thus, need to be translated to their standard forms (Sawaf, 2010; El-taher et al., 2016; Shaalan et al., 2007). Dialects are different from the Modern Standard Arabic (MSA), which is the official language in the Arab world. Studies that investigate dialectal Arabic mainly concentrate on rules and statistic level approaches (Sawaf, 2010; El-taher et al., 2016; Shaalan et al., 2007). While, these approaches need more effort to build the rules, however, the rules can not cover all the words. On the other hand, the static approach produced promising results when applied to some other languages, however, it needs large parallel datasets, dictionary and phrase tables (Mikolov et al., 2013).

Parallel corpus is one of the main components in many machine translation approaches (Xiang et al., 2013). However, this represents a big barrier for low resource languages, such as Arabic and its dialects. There are only few small dialectal datasets, such as the one which was constructed

Proceedings of Recent Advances in Natural Language Processing, pages 52–57,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_008

by Bouamor et al., (2014). Therefore, in this paper we proposed an effective approach that avoids using large parallel corpus and is based on word embedding.

Word embedding is also known also as distributed word representation (Mikolov et al., 2013). It can be implemented using neural networks with the aim of representing words as vectors based on semantic features. Word embedding was used in numerous NLP tasks, such as classification (Rahmawati and Khodra, 2016), language model (Bengio et al., 2003) and sentiment analysis (Altowayan and Tao, 2016). There are many types of word representation methods including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Word2vec and Glove.

In this work, we employed Word2vec to translate the Egyptian dialect text to the Modern Standard Arabic. Monolingual data sets were collected from publicly available sources and a relatively small language pair set was built to train the model.

2 Related work

Unlike English and other international languages, the number of dialectal Arabic NLP studies that involve translation is relatively low. This could be related to a number of reasons that include the existence of various dialects. Arabic and dialectal Arabic could be considered as closely related languages and hence can be translated at the level of word level or character and rules (Sajjad et al., 2013; Durrani et al., 2010; Salloum and Habash, 2011). Previous research on machine translation of dialectal Arabic has focused on normalizing the dialectal word to MSA. Sajjad et al. (2013) have built character level model that attempts to map between dialect and Modern Standard Arabic (MSA) and train on small parallel corpus (Sajjad et al., 2013). The authors used this model to make the translation between Arabic dialect and English more effective (Sajjad et al., 2013). Another proposed approach that can translate dialect to MSA was developed using character level rule and morphological analysis (Sawaf, 2010). Salloum and Habash (2011) proposed a rule based approach that generate the Modern Standard Arabic paraphrases of the low frequency and out-of-vocabulary (OOV) dialect words (Salloum and Habash, 2011). A hybrid system that maps between the Egyptian Arabic and MSA using Egyptian-MSA lexicon and morphological analysis was suggested by Abo Bakr et al. (2008). Zbib et al. (2012) built a language model to translate between dialect and English and trained it on a parallel corpus (Zbib et al., 2012). Furthermore, the Tunisian dialect (TUN) was translated to MSA with deep morphological process based on root and pattern (Hamdi et al., 2013). El-taher et al. (2016) built a model that contains rules, dictionary and language model to understand the context of the Egyptian dialect and translate it to MSA (El-taher et al., 2016). Another method was proposed to translate the Egyptian dialect using rules that are built on top of the Buckwalter Arabic Morphological Analyser (Shaalan et al., 2007).

The above approaches are mostly based on rules that can not cover every word. Also the lack of sufficient parallel data set is still a challenge to translate from any dialect to MSA. To overcome these limitations, we proposed a method that uses word embedding to capture semantic and syntactic features of the word without any rules. The proposed approach emanated from a monolingual data sets rather than parallel corpus. In this study, Word2vec is implemented using Skip-gram and Continuous Bag of Words (CBOW) translation models.

3 Proposed approach

3.1 Word2vec Translation Model

Word2vec was introduced by Mikolov et al. (2013) and it aims to present the words as vectors in low domination space. This model has been successfully applied to a number of NLP tasks such as sentiment analysis, translation and classification (Mikolov et al., 2013). It uses simple neural network (NN) for training and it is considered as prediction based model that can capture linguistic features such as semantic feature (Mikolov et al., 2013; Altowayan and Tao, 2016). There are different parameters that were used for learning NN including the window of the context, the size of the features and negative sample. These parameters help the network to learn representations of the word through training the corpus. Also, it attempts to capture words that are semantically similar between the source and target spaces. Word2vec is based on Skip-gram and Continuous Bag-of-Words (CBOW). The architecture of Skip-gram and CBOW are shown in Figure 1.

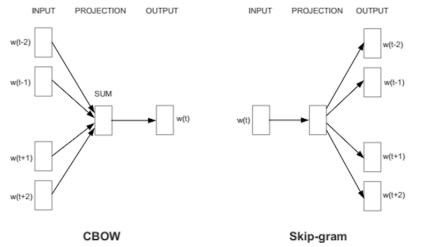

Figure 1: Continuous Bag-of-Word (CBOW) and Skip-gram models (Mikolov et al., 2013)

The model aimed to learn word representation and can be trained on large datasets. The CBOW attempts to predict the target word by combining the distributed representation of its surrounding words whereas, Skip-gram predicts the context by using the distributed representation of the input word (Mikolov et al., 2013). Also, there are two basic training objectives; hierarchical softmax and negative sampling. The Skip gram and CBOW use neural network (NN) to predict the neighbouring words by learning the word representation (Rong, 2014; Enríquez et al., 2016). Moreover, CBOW is fast and appropriate for large corpus while Skip-gram can be trained on small monolingual data sets.

$$skip-gram(D) = \frac{1}{T}\sum_{t=1}^{T}[\sum_{j=-k}^{k}\log p(w_{t+j}|w_t)] \quad (1)$$

$$CBOW(D) = \frac{1}{T}\sum_{t=1}^{T}[\sum_{j=k}^{t-k}\log p(w_j|w_{j+t})] \quad (2)$$

Suppose *D*= $w_1,..,w_t$ where k is the size of training context, equation (1) aims to maximize the average of log probability to predict the context words w_{t+j} based on the current word w_t (Mikolov et al., 2013). Whereas equation (2) computes the log probability of the target word w_i based on the surrounding words in the context (Chen et al., 2015).

Linear mapping can capture similar vectors from the source and target languages. Therefore, this mapping can learn the word translation matrix between different languages.

Translation matrix is based on a set of word pairs and their vectors. Consider $\{x_i, z_i\}_{i=1}^{n}$, where x_i and z_i are word vectors. These vectors have different dimensions (d_1, d_2), $x_i \in R_{d_1}$ is the representation of word *i* in source language and $z_i \in R_{d_2}$ is the vector of the translation (Mikolov et al., 2013). The translation matrix can be learned by equation (3)(Mikolov et al., 2013).

$$min\Sigma_{i=1}^{n}||W_{x_i-z_i}||^2 \quad (3)$$

3.2 Language challenge

The Arabic language is considered as one of the six official languages of the United Nations (Aljlayl and Frieder, 2002; Ibrahim et al., 2015). It is spoken by 300 million people globally and considered to be a morphologically rich language (Cheriet, 2007; Aljlayl and Frieder, 2002). It has different structures from English and other languages. The Arabic language contains 28 letters and is written from right to left. There are two types of Arabic, Modern Standard Arabic (MSA) which is the formal Arabic language used in newspapers and books; and spoken varieties or Arabic dialect (DA), which is the language that is used in daily life and in social media (El-taher et al., 2016). There are different DAs such as Egyptian, Yemeni, Gulf, Iraqi and Levantine (Sajjad et al., 2013). However, these dialects are different from each other depending on the geographical distribution (Habash, 2010).

Arabic Natural Language Processing faces many challenges because Arabic is a morphologically rich language (Salloum and Habash, 2011). Arabic dialects are different from MSA and different from each other. Also, there are different features between dialects and Arabic as there is no rule for written set of grammar. For example, variation might be appeard orthographically, lexically and morphologically (Habash, 2010). In dialects there is no standard orthography which lead every dialects to spell same word in different ways, for instance (ميه ، مويه ، ماء) for water. Also, the ambiguity due to using diacritical marks which called Tashkiil in Arabic, this changes the meaning for the same word for example (شَعَرَ ، شَعَر ، شِعْر) for hair, feel and poetry respectively, where the diacritical marks make these words that are formed using the same letters having different meanings. Another feature is misspelling in dialect as they spell differently in MSA; for example, the word gold can be written as (ذهب) in MSA and as (دهب) in EGY. Although, these variations between Arabic and the various dialects, there is also similar semantic featuresas a result of similarity between them.

3.3 Pre-processing (Normalization)

Pre-processing is recognized as an essential step for a number of NLP tasks. Text normalization is one type of text pre-processing, which is defined as a process of transforming the non-standard words to their standard forms. For example, 2morrw should be transformed into tomorrow. Text normalization plays a major role in a number of Arabic Natural Language Processing tasks, such as information retrieval which included sentiment analysis, summarization, keywords, and topic detection. Arabic normalization may include deleting the diacritical marks to reduce the ambiguity. Consequently, we appliedd some normalization steps to clean our data sets and prepared them for the translation process. These steps included:

- Tokenization.
- Delete any diacritics from Arabic letter (Tshkula).
- Replace إ، آ، أ with ا, replace ة with ه , replace ؤ with وand replace ئ with ى.
- Remove non-Arabic words and punctuation marks (?, !).

4 Experiment and Result

4.1 Building monolingual Corpus

As a basic requirement for machine translation and other NLP tasks, data sets are needed to implement and validate proposed models. Both of parallel data and dictionaries are important for translation tasks. However, Arabic lacks sufficient parallel corpus. We could only find relatively small parallel corpus. Also, unlike some other languages there is no available parallel dictionary for Arabic dialects. Thus, we firstly built monolingual corpus for both the Egyptian and standard Arabic from Wikipedia and different resources that are publicly available. In this experiment we used four datasets for target language; MSA-EGY Wikipedia, bbc-arabic , osac-utf-8 corpus and lastly, cnn-arabic , Table 1 shows details for the data sets. Secondly, for the Egyptian dialect we had to construct a database with a reasonable size that incorporated the parallel data described in (Bouamor et al., 2014).

Table 1: Arabic and Egyptian data sets

Name	**size**
MSA-EGY Wikipedia	862MB
cnn-arabic	24MB
bbc-arabic	21MB
osac-uft8	178MB
MSA*	89KB
Egyptian*	86KM
Egyptian(own data)	143KM
parallel dictionary	375KM

4.2 Experiment

The following word2vec processes are used to translate the Egyptian dialect to the Modern Standard Arabic. Firstly, we normalized both the Arabic and the Egyptian data sets by using the steps that were mentioned in the pre-processing section. Secondly, two separate word vector models for target and source languages were built using CBOW and Skip-gram models. These models were applied on the data sets that were described in the previous section. The model parameters were set as 100 for the size of features, 5 for window size and 2 for minimum count which mean deleting any word that appear less than two times. Then, translation matrix was trained on the Arabic-Egyptian language pairs. In order to find semantic words translation based on the context, the model was trained on the monolingual data sets. Finally, testing was done in four-fold-cross-validation, i.e., 75% for training and 25% for testing.

4.3 Results

Monolingual data sets were used to evaluate our proposed model using Top@1 and Top@5 accuracy scores and a four-fold-cross-validation approach. As the translation was based on context, the predicted words are expected to be semantically and syntactically related to target words. Figure 2 shows the average accuracy of CBOW for Top@5 and Top@1 when applied to the four data sets of osac-utf8, Wikipedia, bbc-arabic and cnn-arabic, while Figure 3 shows the accuracy of Skip-gram model. The two figures show that CBOW produces better and more consistent results than Skip-gram. More specifically, the CBOW accuracy for all four datasets ranged between 77% and 81% and between 63% and 73%, for the Top@5 and Top@1 scores respectively. On the other hand, apart from the bbc-arabic data set, the Skip-

gram was not found to achieve good result for the remaining three data sets (see Table 2). These results pointed out that the CBOW monolingual model was able to capture better semantically related words than Skip-gram. Also, the training time of CBOW is found to be faster than that of Skip-gram. Below are some examples of words and their translation as derived from CBOW and Skip-gram.

- عندهم [لديهم، لديه، لدينا، لديه، ثمه]
- هيبقى [سيبقى، سيظهر، ياخذ، للتاكيد، لايزال]
- إتنازلت [تخلت، استقلت، تخلي، تتنازل، تنازلت]
- ويعيش [ويعيش، يعيش، يتربى، فيعيش، فعاش]

English translation :

- They have[they have, he has, we have, he has, there is]
- Will remain[will remain, will appear, takes, for confirmation, still]
- Waived[waived, abandoned, resigned, waive, give up]
- Live[live, live, grow up, lived, lived]

As presented in the list above the words translated without any rules and some words translated based on the context. Even though, in some cases some words were not correctly translated, they still produced semantically related words e.g. القلق [النوم، والتوتر، والترنح ، الدوخه، الترهل] which means in English Anxiety [sleep, tension, grogginess, dizziness, sag] which are all semantically related to the word anxiety.

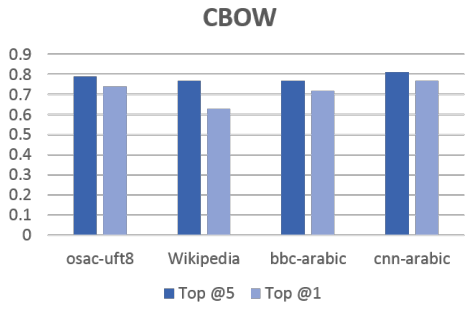

Figure 2: Top@5 and Top@1 for CBOW model.

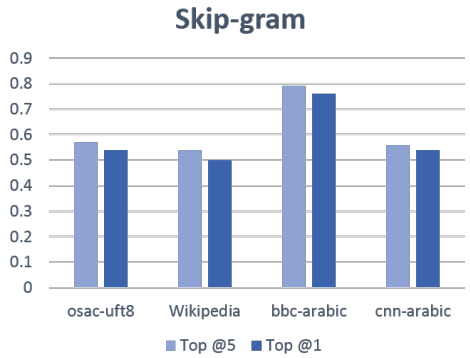

Figure 3: Top@5 and Top@1 for Skip-gram model

Table 2: The average of four-fold-cross-validation of all data sets

Model Name	Data set	Top@5	Top@1
CBOW	cnn-arabic	0.81	0.77
	bbc-arabic	0.77	0.72
	osac-utf8	0.79	0.74
	Wikipedia	0.77	0.63
Skip-gram	cnn-arabic	0.56	0.54
	bbc-arabic	0.79	0.76
	osac-utf8	0.57	0.54
	Wikipedia	0.54	0.50

5 Conclusion

Word embedding is a powerful approach in NLP. In this paper, word2vec was introduced to translate the Egyptian dialect to the Modern Standard Arabi. This approach solves the problem of parallel data as we can train the model on monolingual data. Word2vec has also shown that it can capture semantic features between MSA and EGY without any rules. Even though the model was only tested on small data set, it is expected to also perform well on large data sets. In future work, we plan to investigate the effect of other features at character level and other morphological features.

References

Fahad Albogamy and Allan Ramsay. 2015. POS Tagging for Arabic Tweets. In *RANLP*. pages 1–8.

Mohammed Aljlayl and Ophir Frieder. 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. In *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, pages 340–347.

A. A. Altowayan and L. Tao. 2016. Word embeddings for Arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)*. pages 3820–3825. https://doi.org/10.1109/BigData.2016.7841054.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research* 3(Feb):1137–1155.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *LREC*. pages 1240–1245.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint Learning of Character and Word Embeddings. In *IJCAI*. pages 1236–1242.

Mohamed Cheriet. 2007. Strategies for visual arabic handwriting recognition: issues and case study. In *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*. IEEE, pages 1–6.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 465–474.

Fatma El-zahraa El-taher, Alaa Aldin Hammouda, and Salah Abdel-Mageid. 2016. Automation of understanding textual contents in social networks. In *Selected Topics in Mobile & Wireless Networking (MoWNeT), 2016 International Conference on*. IEEE, pages 1–7.

Fernando Enríquez, José A Troyano, and Tomás López-Solaz. 2016. An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications* 66:1–6.

Nizar Y Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3(1):1–187.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. The effects of factorizing root and pattern mapping in bidirectional Tunisian-standard Arabic machine translation. In *MT Summit 2013*. pages pas–d.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 368–378.

Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. 2015. MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis. In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*. IEEE, pages 353–358.

Chen Li and Yang Liu. 2012. Normalization of Text Messages Using Character-and Phone-based Machine Translation Approaches. In *INTERSPEECH*. pages 2330–2333.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Dyah Rahmawati and Masayu Leylia Khodra. 2016. Word2vec semantic representation in multilabel classification for Indonesian news article. In *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On*. IEEE, pages 1–6.

Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* .

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating Dialectal Arabic to English. In *ACL (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*. Association for Computational Linguistics, pages 10–21.

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*.

Khaled Shaalan, Hitham Bakr, and Ibrahim Ziedan. 2007. Transferring egyptian colloquial dialect into modern standard arabic. In *International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria*. pages 525–529.

Lu Xiang, Yu Zhou, and Chengqing Zong. 2013. An Efficient Framework to Extract Parallel Units from Comparable Data. In *Natural Language Processing and Chinese Computing*, Springer, pages 151–163.

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*. Association for Computational Linguistics, pages 49–59.