Automatic Identification of Arabic Language Varieties and Dialects in Social Media

Fatiha Sadat
University of Quebec in Montreal, 201 President Kennedy, Montreal, QC, Canada
sadat.fatiha@uqam.ca

Farnazeh Kazemi
NLP Technologies Inc.
52 Le Royer Street W., Montreal, QC, Canada
kazemi@nlptechnologies.ca

Atefeh Farzindar
NLP Technologies Inc.
52 Le Royer Street W., Montreal, QC, Canada
farzindar@nlptechnologies.ca

Abstract

Modern Standard Arabic (MSA) is the formal language in most Arabic countries. Arabic Dialects (AD) or daily language differs from MSA especially in social media communication. However, most Arabic social media texts have mixed forms and many variations especially between MSA and AD. This paper aims to bridge the gap between MSA and AD by providing a framework for AD classification using probabilistic models across social media datasets. We present a set of experiments using the character n-gram Markov language model and Naive Bayes classifiers with detailed examination of what models perform best under different conditions in social media context. Experimental results show that Naive Bayes classifier based on character bi-gram model can identify the 18 different Arabic dialects with a considerable overall accuracy of 98%.

1 Introduction

Arabic is a morphologically rich and complex language, which presents significant challenges for natural language processing and its applications. It is the official language in 22 countries spoken by more than 350 million people around the world[1]. Moreover, the Arabic language exists in a state of diglossia where the standard form of the language, Modern Standard Arabic (MSA) and the regional dialects (AD) live side-by-side and are closely related (Elfardy and Diab, 2013). Arabic has more than 22 dialects; some countries share the same dialect, while many dialects may exist alongside MSA within the same Arab country.

Modern Standard Arabic (MSA) is the written form of Arabic used mostly in education and scripted speech; it is also the formal communication language. Spoken Arabic is often referred to as colloquial Arabic, dialects, or vernaculars. Thus. Arabic dialects (AD) or colloquial languages are spoken varieties of Arabic and the daily language of several people. Arabic dialects and MSA share a considerable number of semantic, syntactic, morphological and lexical features; however, these features have many differences (Al-Sabbagh and Girju, 2013).

Recently, considerable interest was given to Arabic dialects and the written varieties of Arabic found on social networking sites such as chats, micro-blog, blog and forums, which is the target research of sentiment analysis, opinion mining, machine translation, etc.

Social media poses three major computational challenges, dubbed by Gartner the 3Vs of big data: *volume, velocity, and variety*[2]. NLP methods, in particular, face further difficulties arising from the short, noisy, and strongly contextualised nature of social media. In order to address the 3Vs of social media, new language technologies have emerged, such as the identification and definition of users' language varieties and the translation to a different language, than the source.

[1] http://en.wikipedia.org/wiki/Geographic_distribution_of_Arabic#Population
[2] http://en.wikipedia.org/wiki/Big_data

Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 22–27,
Dublin, Ireland, August 24 2014.

Dialect identification is essential and considered the first prepossessing component for any natural language application dealing with Arabic and its variation such as machine translation, information retrieval for social media, sentiments analysis, opinion extraction, etc.

Herein, we present our effort on a part of the ASMAT project (*Arabic Social Media Analysis Tools*), which aims at creating tools for analyzing social media in Arabic. This project paves the way for end user targets (like machine translation and sentiment analysis) through pre-processing and normalization. There are, however, still many challenges to be faced.

This paper presents a first-step towards the ultimate goal of identifying and defining languages and dialects within the social media text. This paper is organized as follows: Section 2 presents related work. Sections 3 and 4 describe the probabilistic approach based on the character n-gram Markov language model and Naive Bayes classifier. Section 5 presents the data set, the several conducted experiments and their results. Conclusions and future work are presented in Section 6.

2 Related Work

There have been several works on Arabic Natural Language Processing (NLP). However, most traditional techniques have focused on MSA, since it is understood across a wide spectrum of audience in the Arab world and is widely used in the spoken and written media. Few works relate the processing of dialectal Arabic that is different from processing MSA. First, dialects leverage different subsets of MSA vocabulary, introduce different new vocabulary that are more based on the geographical location and culture, exhibit distinct grammatical rules, and adds new morphologies to the words. The gap between MSA and Arabic dialects has affected morphology, word order, and vocabulary (Kirchhoff and Vergyri, 2004). Almeman and Lee (2013) have shown in their work that only 10% of words (unigram) share between MSA and dialects.

Second, one of the challenges for Arabic NLP applications is the mixture usage of both AD and MSA within the same text in social media context. Recently, research groups have started focusing on dialects. For instance, Columbia University provides a morphological analyzer (MAGAED) for Levantine verbs and assumes the input is non-noisy and purely Levantine (Habash and Rambow, 2006).

Dialect Identification task has the same nature of language identification (LI) task. LI systems achieved high accuracy even with short texts (Baldwin and Lui, 2010), (Cavnar and Trenkle, 1994), (Joachims, 1998), (Kikui,1996); however, the challenge still exists when the document contains a mixture of different languages, which is actually the case for the task of dialect identification, where text is a mixture of MSA and dialects, and the dialects share a considerable amount of vocabularies. Biadsy and al. (2009) present a system that identifies dialectal words in speech and their dialect of origin through the acoustic signals. Salloum and Habash (2011) tackle the problem of AD to English Machine Translation (MT) by pivoting through MSA. The authors present a system that applies transfer rules from AD to MSA then uses state of the art MSA to English MT system. Habash and al. (2012) present CODA, a Conventional Orthography for Dialectal Arabic that aims to standardize the orthography of all the variants of AD while Dasigi and Diab (2011) present an unsupervised clustering approach to identify orthographic variants in AD.

Recently, Elfardy and Diab (Elfardy and al., 2013) introduced a supervised approach for performing sentence level dialect identification between Modern Standard Arabic and Egyptian Dialectal Arabic. The system achieved an accuracy of 85.5% on an Arabic online-commentary dataset outperforming a previously proposed approach achieving 80.9% and reflecting a significant gain over a majority baseline of 51.9% and two strong baseline systems of 78.5% and 80.4%, respectively (Elfardy and Diab, 2012).

Our proposed approach for dialect identification focuses on character-based n-gram Markov language models and Naive Bayes classifiers.

Character n-gram model is well suited for language identification and dialect identification tasks that have many languages and/or dialects, little training data and short test samples.

One of the main reasons to use a character-based model is that most of the variation between dialects, is based on affixation, which can be extracted easily by the language model, though also there are word-based features which can be detected by lexicons.

3 N-Gram Markov Language Model

There are two popular techniques for language identification. The first approach is based on popular words or stop-words for each language, which score the text based on these words (Gotti and al., 2013). The second approach is more statistical oriented. This approach is based on n-gram model (Cavnar and Trenkle, 1994), Hidden Markov model (Dunning, 1994) and support vector machine (Joachims, 1998).

A language model is one of the main components in many NLP tools and applications. Thus, lot of efforts have been spent for developing and improving features of the language models.

Our proposed approach uses the Markov model to calculate the probability that an input text is derived from a given language model built from training data (Dunning, 1994). This model enables the computation of the probability *P(S)* or likelihood, of a sentence *S,* by using the following chain formula in equation 1:

$$P(w_1,w_2,...,w_n) = P(w_1) \prod_{i=2}^{n} P(w_i|w_1 ... w_{i-1}) \qquad (1)$$

Where, the sequence (w_1, …, w_n) represents the sequence of characters in a sentence *S*. $P(w_i \mid w_1, \ldots, w_{i-1})$ represents the probability of the character w_i given the sequence w_1, …, w_{i-1}.

Generally, the related approach that determines the probability of a word sequence is not very helpful because of its computational cost that is considered as very expensive.

Markov models assume that we can predict the probability of some future unit without looking too far into the past. So we could apply the Markov assumption to the above chain probability in Formula 1, by looking to zero character (uni-gram), one character (bi-gram), two characters (tri-gram).

The intuition behind using n-gram models in a dialect identification task is related to the variation in the affixations that are attached to words, which can be detected by bi-gram or tri-gram models.

4 Naïve Bayed Classifier

Naive Bayes classifier is a simple and effective probabilistic learning algorithm. A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[3].

In text classification, this classifier assigns the most likely category or class to a given document *d* from a set of pre-define N classes as c_1, c_2, ..., c_N. the classification function *f* maps a document to a category *(f: D ➔ C)* by maximizing the probability of the following equation (Peng and Schuurmans, 2003):

$$P(c|d) = \frac{P(c) \times P(d \mid c)}{P(d)} \qquad (7)$$

Where, *d* and *c* denote each the document and the category, respectively. In text classification a document *d* can be represented by a vector of *T attribute* $d=(t_1,t_2, ..., t_T)$.

Assuming that all attribute t_i are independent given the category *c*, then we can calculate *p(d|c)* with the following equation:

$$\underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} P(c) \times \prod_{i=1}^{T} P(t_i|c) \qquad (8)$$

The attribute term t_i can be a vocabulary term, local n-gram, word average length, or global syntactic and semantic properties (Peng and Schuurmans, 2003).

5 Experiments and Results

We have carried out two sets of experiments. The first set of experiments uses the character n-gram language model, while the second one uses the Naive Bayes classifier. The developed system identifies Arabic dialects using character n-gram models where the probability of each (Uni-gram, Bi-gram and Tri-gram) is calculated based on the training data within social media context.

[3]http://en.wikipedia.org/wiki/Naive_Bayes_classifier

5.1 Data

The System has been trained and tested using a data set collected from blogs and forums of different countries with Arabic as an official language. We have considered each regional language or dialect as belonging to one Arab country, although in reality a country most of the time may have several dialects. Moreover, there is a possible division of regional language within the six regional groups, as follows: Egyptian, Levantine, Gulf, Iraqi, Maghrebi and others (Zaidan and Callison-Burch, 2012). The different group divisions with their involved countries are defined as follows:

- *Egyptian:* Egypt;
- *Iraqi:* Iraq;
- *Gulf*: Bahrein, Emirates, Kuwait, Qatar, Oman and Saudi Arabia;
- Maghrebi: Algeria, Tunisia, Morocco, Libya, Mauritania;
- Levantine: Jordan, Lebanon, Palestine, Syria;
- Others: Sudan.

Moreover, there might be many other possible sub-divisions in one division, especially in the large region such as the Maghrebi. We used a data set that consists on the crowd source of social media texts such as forums and blogs. This set of data was manually collected and constructed using several (around eighteen) forums sites in Arabic. The collected texts were manually segmented to coherent sentences or paragraphs. For each dialect, sentences were saved in XML format with additional information such as sequence number, country, date, and the link. Table 1 shows some statistics about the collected text such as the total number of sentences or paragraph and number of words for each dialect. For each dialect 100 sentences were selected randomly for test purposes and were excluded from the training data. Moreover, statistics on the data set for each group of countries can be constructed, following the data set of Table 1.

Country	***#sentences***	***#words***
Egypt	7 203	72 784
Bahrain	3 536	36 006
Emirates	4 405	43 868
Kuwait	3 318	44 811
Oman	4 814	77 018
Qatar	2 524	22 112
Saudi Arabia	9 882	82 206
Jordon	1 944	18 046
Lebanon	3 569	26 455
Palestine	316	3 961
Syria	3 459	43 226
Algeria	731	10 378
Libya	370	5 300
Mauritania	2 793	62 694
Morocco	2 335	30 107
Tunisia	3 843	18 199
Iraq	1 042	13 675
Sudan	5 775	28 368

Table 1: Statistics about the dataset for each country

5.2 Results and Discussion

We have carried out three different experiments using uni-gram, bi-gram and tri-gram character for each experiment base on either the Markov language model or the Naive Bayes classifier. These different experiments show how character distribution (uni-gram) or the affixes of size 2 or 3 (bi-gram or tri-gram) help distinguish between Arabic dialects. The set of experiments were conducted on 18 dialects representing 18 countries. Furthermore, we conducted the experiments on six groups of Arabic dialects, which represent six areas as described in the earlier section.

For evaluation purposes, we considered the accuracy as a proportion of true identified test data and the *F*.Measure as a balanced mean between precision and recall. Our conducted experiments showed that the character-based uni-gram distribution helps the identification of two dialects, the Mauritanian and the Moroccan with an overall F-measure of 60% and an overall accuracy of 96%. Furthermore, the bi-gram distribution of two characters affix helps recognize four dialects, the Mauritanian, Moroccan, Tunisian and Qatari, with an overall F-measure of 70% and overall accuracy of 97%.

Last, the tri-gram distribution of three characters affix helps recognize four dialects, the Mauritanian, Tunisian, Qatari and Kuwaiti, with an overall F-measure of 73% and an overall accuracy of 98%. Our comparative results show that the character-based tri-gram and bi-gram distributions have performed better than the uni-gram distribution for most dialects. Overall, for eighteen dialects, the bi-gram model performed better than other models (uni-gram and tri-gram).

Since many dialects are related to a region, and these Arabic dialects are approximately similar, we also consider the accuracy of dialects group. Again, the bi-gram and tri-gram character Markov language model performed almost same, although the F-Measure of bi-gram model for all dialect groups is higher than tri-gram model except for the Egyptian dialect. Therefore, in average for all dialects, the character-based bi-gram language model performs better than the character-based uni-gram and tri-gram models.

Our results show that the Naive Bayes classifiers based on character uni-gram, bi-gram and tri-gram have better results than the previous character-based uni-gram, bi-gram and tri-gram Markov language models, respectively. An overall F-measure of 72% and an accuracy of 97% were noticed for the eighteen Arabic dialects. Furthermore, the Naive Bayes classifier that is based on a bi-gram model has an overall F-measure of 80% and an accuracy of 98%, except for the Palestinian dialect because of the small size of data. The Naive Bayes classifier based on the tri-gram model showed an overall F-measure of 78% and an accuracy of 98% except for the Palestinian and Bahrain dialects. This classifier could not distinguish between Bahrain and Emirati dialects because of the similarities on their three affixes. In addition, the naive Bayes classifier based on a character bi-gram performed better than the classifier based on a character tri-gram. Also, the accuracy of dialect groups for the Naive Bayes classifier based on character bi-gram model yielded better results than the two other models (uni-gram and tri-gram).

6 Conclusion

In this study, we presented a comparative study on dialect identification of Arabic language using social media texts; which is considered as a very hard and challenging task. We studied the impact of the character n-gram Markov models and the Naive Bayes classifiers using three n-gram models, uni-gram, bi-gram and tri-gram. Our results showed that the Naive Bayes classifier performs better than the character n-gram Markov model for most Arabic dialects. Furthermore, the Naive Bayes classifier based on character bi-gram model was more accurate than other classifiers that are based on character uni-gram and tri-gram. Last, our study showed that the six Arabic dialect groups could be distinguished using the Naive Bayes classifier based on character n-gram model with a very good performance.

As for future work, it would be interesting to explore the impact of the number of dialects or languages on a classifier. Also, it would be interesting to explore the influence of size of training and test set for both character n-gram Markov model and Naive Bayes classifier based on character n-gram model. We are planning to use more social media data from Twitter or Facebook in order to estimate the accuracy of these two models in the identification of the dialect and the language. Another extension to this work is to study a hybrid model for dialect identification involving character-based and word-based models. Finally, what we presented in this draft is a preliminary research on exploiting social media corpora for Arabic in order to analyze them and exploit them for NLP applications. Further extensions to this research include the translation of social media data to other languages and dialects, within the scope of the ASMAT project.

Reference

Al-Sabbagh R. and Girju R. 2013. Yadac : Yet another dialectal arabic corpus. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 2012.

Almeman K. and Lee M. 2013. Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In Communications, Signal Processing, and their Applications (ICCSPA), 2013.

Baldwin T. and Lui M. 2010. Language identification: The long and the short of the matter. In Human Language Technologies: The 2010 Annual Conference of theNorth American Chapter of the Association for Computational Linguistics, HLT'10, pages 229{237, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Biadsy F., Hirschberg J., and Habash N. 2009. Spoken arabic dialect identification using phonotactic modeling. In Proceedings of the Workshop on Computational Approaches to Semitic Languages at the meeting of the European Association for Computational Linguistics (EACL), Athens, Greece.

Cavnar W. B., Trenkle J. M. 1994. N-gram-based text categorization. 1994. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Ann Arbor MI, 48113(2):161{175, 1994.

Dasigi P. and Diab M. 2011. Codact: Towards identifying orthographic variants in dialectal arabic. In Proceedings of the 5th International Joint Conference on Natural Language Processing (ICJNLP), Chiangmai, Thailand, 2011.

Dunning T. 1994. Statistical identification of languages. Citeseer, 1994.

Elfardy H. and Diab M. 2012. Simplified guidelines for the creation of large scale dialectal arabic annotations. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Elfardy H. and Diab M. 2013. Sentence-Level Dialect Identification in Arabic, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Sofia, Bulgaria. 2013.

Elfardy H., Al-Badrashiny M., Elfardy M. and Diab M. 2013. Sentence Level Dialect Identification in Arabic. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 456–461, Sofia, Bulgaria, August 4-9 2013.

Gotti F., Langlais P., and Farzindar A. 2013. Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In Proceedings of the Workshop on Language Analysis in Social Media, Atlanta, Georgia, June 2013. Association for Computational Linguistics, Association for Computational Linguistics.

Habash N. and Rambow O. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics , pages 681–688, Sydney, Australia, July. Association for Computational Linguistics.

Habash N., Diab M., and Rabmow O. 2012. Conventional orthography for dialectal arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey, 2012.

Joachims T. 1998. Text categorization with support vector machines: Learning with many relevant features. Springer, 1998.

Kikui G.-i. 1996. Identifying, the coding system and language, of on-line documents on the internet. In Proceedings of the 16th Conference on Computational Linguistics - Volume 2, COLING '96, pages 652{657, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

Kirchhoff K. and Vergyri D. 2004. Cross-dialectal acoustic data sharing for arabic speech recognition. In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, volume 1, pages I-765. IEEE, 2004.

Peng F. and Schuurmans D. 2003. Combining naive bayes and n-gram language models for text classi_cation. In Advances in Information Retrieval, pages 335{350. Springer, 2003.

Salloum W. and Habash N. 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.

Suliman A. F. 2008. Automatic Identification of Arabic Dialects USING Hidden Markov Models. Doctoral Dissertation, University of Pittsburgh. 2008.

Zaidan O. F. and Callison-Burch C. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In Proceedings of ACL, pages 37–41, 2011.

Zaidan O. F. and Callison-Burch C. 2012. Arabic dialect identi_cation. volume 1, Microsoft Research, 2012.