*Topic Modeling of Phonetic Latin-Spelled Arabic for the Relative Analysis of Genre-Dependent and Dialect-Dependent Variation **

Ali SAKR, Mark HASEGAWA-JOHNSON

[1]LIDILEM / Université Stendhal - Grenoble III
Domaine Universitaire - 1180, avenue centrale F-38400 Saint Martin d'Hères, France
[2]Institut CEA, LIST, DIASI / Laboratoire Vision et Ingénierie des Contenus
CEA Saclay – Nano-INNOV Bâtiment 861 – PC 173 F-91191 Gif-sur-Yvette Cedex, France

alisakr1@illinois.edu , jhasegaw@illinois.edu

Abstract: *We demonstrate a data collection and analysis system that can be used to analyze the relative contributions of dialect dependent variation in the lexical of speech-like Arabic text. We utilize Latent Dirichlet Allocation (LDA), a generative Probabilistic modeling method, to analyze a phonetic Latin Spelled Arabic online chat corpus. The corpus produces different word choices and word relations based on Dialect, which can therefore aid in producing written forms of Arabic Dialects despite the large difference between Standard Written Arabic and the many Arabic Dialects.*

Résumé : *Nous présenterons un système de collecte et d'analyse de données éventuellement utilisé pour analyser les contributions relatives des variations dépendantes au dialecte dans la sphère lexicale d'un texte semblable à l'écriture arabe. De ce fait, nous aurons recours à l'allocation de Dirichlet latente (LDA), une méthode de modélisation générative probabiliste afin d'analyser la phonétique des termes arabes écrits en caractère latin extraits d'un corpus de discussion en ligne. Ce corpus produit différents choix de mots et différentes relations conceptuelles basée sur le dialecte et qui par conséquent contribue à la reproduction graphique des termes arabes issus du dialecte malgré la large distinction existante entre l'arabe écrit standard et les nombreux dialectes arabes.*

Keywords: *Topic Modeling, phonetic Latin-Spelled Arabic, LDA, Arabic online chat corpus analysis.*
Mots clés *: Modélisation thématique, phonétique des termes arabes écrits en caractère latin, LDA, analyse d'un corpus de discussion en ligne.*

**Modélisation thématique de la phonétique des termes arabes écrits en caractère latin pour une analyse relative des variations dépendantes au genre et celles dépendantes au dialecte*

© Revue RIST, vol. 20, n° 2, p.69-77

1 Introduction

Currently, Arabic Speech to text transcription faces a problem that speech-to-text transcriptions of other languages do not. The difference between the spoken Arabic and the written Arabic is so vast that it is difficult to map spoken Arabic to corresponding written Arabic. Additionally, there is a wide array of Arabic Dialects. The contemporary attempt to solve this problem is through massive data collection of speech from "Talk shows, debates, and interactive [television and radio] programs," [1]. In this paper we attempt to use a new source of data for Arabic text. We look at online conversation, specifically YouTube comments on Arabic videos written in Latin Spelled phonetic Arabic. Our data analysis demonstrates systematic variation in word choice as a function of both chat genre and national dialect, suggesting that data of this kind might be used to bridge the gap between spoken Arabic and Modern Arabic, e.g. for the purpose of developing better Arabic-language automatic speech recognition.

Based on the research completed to write this paper, this is perhaps one of, if not, the first instances in which online chat data is collected for the purpose of aiding speech-to-text transcription. However, numerous other studies have documented significant textual features of online chat that differ from formal writing within the same language [11].

Although online chat may relate to spoken Arabic more than Standard written Arabic; it concurrently has its own set of problems. Being an unofficial language, there is no uniform spelling and no spell checker to run the data through. This is something that can be developed as data collection of online Latin Spelled Phonetic Arabic continues. The lack of uniform spelling can cause problems in topic modeling. The lack of uniform spelling can distort word frequency of a given topic, since the frequencies of a given word will be broken into several different spellings. A system for determining a most likely spelling of a given term could greatly enhance analysis of a text corpus with no uniform spelling. Potential Solutions to this problem are discussed in the Results and Analysis section of the paper.

2 Background

Latent Dirichlet Allocation (LDA) is perhaps the most common topic modeling method in information retrieval. LDA was developed in 2003 and has since been used as "a springboard for many other topic models" [8]. LDA allows terms to be categorized into groups or topics. In addition to analysis of text, LDA, as well as more advanced topics models that are LDA based, has also been used for analysis of genomic data [9] and of audio data [10].

Latent Dirichlet Allocation is a "generative probabilistic model for collections of discrete data," [2]. LDA uses hierarchical Bayesian modeling to determine the likelihood of a topic distribution, which then can allow analysis of topic distributions in documents given word frequency differences between different topics, and documents.

Since there are numerous topics that create the topic distribution, the probability distribution is of a multinomial vector (θ) in which each variable (θ_i) of the vector (θ) is such that, $0 \leq \theta_i \leq 1, \sum_{i=1}^{k} \theta_i = 1$, where k is the number of variables or dimensions in the vector θ [2]. The probability of the vector θ is the product of the probability of each of the vector's dimension's (θ_i) given a vector of occurrences (α) of each θ_i, is determined via Dirichlet Distribution.

The formula for the probability of the vector of θ, given α is $P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\ \theta_1^{\alpha_1-1} \ldots\ldots \theta_k^{\alpha_k-1}$ [2]. The calculation is the inverse of the beta distribution (where the Beta distribution = $(\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)})$ ^-1), multiplied by the likelihood of the exact vector α of occurrences given the topic distribution θ. The inverse of the Beta distribution serves to count for all the possible combinations of the vector α fitting the topic distribution θ. This is similar to how $\binom{n}{k} = \frac{n!}{k!(n-k!)}$ functions as the number of unordered combinations in binomial probabilities. The total ($P(\theta|\alpha)$) is the probability that the topic distribution θ is appropriate for the given data represented by vector α, ($P(\theta|\alpha)$).

Another input parameter to the LDA model is multinomial lexicon probabilities matrix β. Beta (β) is a matrix of dimensions k X V, where k is the number of topics, and therefore also the number of dimensions of the vector θ. V is the number of unique words in all of the topics combined. LDA uses the word frequencies in each of the documents of the corpus to estimate the word distributions in each of k different topics and then estimate topic distributions in each of the documents. The topic distribution estimates of each document are represented by a vector θ. Supposing we have a set of N topics and a set of N words, the equation for the joint distribution of topics, words, and the topic distribution θ is the following[2].

$$P(\theta, z, \boldsymbol{w}|\alpha, \beta) = P(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)P(w_n|z_n, \beta)$$

The **w** in bold denotes a document, while a w refers simply to a word.

The topic distribution θ, given that θ ~ Dir(α) (meaning θ is a Dirichlet distribution parameterized by α), is a continuous multinomial vector. The set of topics is of course a discrete set. So, by integrating over θ and summating over the topics, we get the Marginal distribution of a given document.

$$P(\boldsymbol{w}|\alpha, \beta) = \int P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{k=1}^{x} p(z_n|\theta)\ P(w_n|z_n, \beta)d\theta\right)$$

Each document **w**, from a set of documents (a corpus), has a marginal probability defined in the sentence above. These marginal probabilities can by multiplied together to produce a probability, of the entire set of documents (a corpus *D containing*). The corpus probability $p(D|\alpha, \beta)$ is the following equation [2].

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int P(\theta|\alpha)(\prod_{n=1}^{N}\sum_{k=1}^{x} p(z_n|\theta)\ P(w_n|z_n, \beta)d\theta_d)$$

Latent Dirichlet Allocation provides information retrieval researchers an effective modeling of topics as well as a methodology for determining inter-document relation based on topic distribution (which is of course determined from the word frequencies of each document as well as the matrix β and the term frequencies of each topic that said matrix provides).

Latent Dirichlet Allocation has proven to be an effective technique for topic modeling of inter-document relation. Other techniques utilizing Dirichlet priors, such as Latent

Dirichlet Language Modeling (LDLM), have improved accuracy of classifying and relating word sequences [3]. For this reason, LDA is an appropriate technique for language modeling for speech recognition speech-to-text transcription. However, LDA is not any more appropriate than other methods that utilize Dirichlet priors such as LDLM or Sentence-based Latent Dirichlet Allocation (SLDA) [4].

3 Experimental methods

In order to have our data ran successfully through the LDA Expectation Maximization (EM) training algorithm, we first manipulate character classes to collect Arabic chat words and format each document into a vector of lexicon term counts. In addition to letters, numbers are included in the character class because numbers are used in Arabic online chat. Naturally, since Arabic doesn't have an alphabet that maps one-to-one to the English alphabet or any Latin alphabet, numbers are part the "Arabic Chat Alphabet " that has been informally adopted by many online Arabic Language Communities [5]. El-Mahdy's 2011 publication provides a table of this Alphabet, which El-Mahdy terms the "Arabic Chat Alphabet."

There is a great deal of stylistic variability in the spellings of the Arabic Chat Alphabet. In attempt to normalize some of the spelling variability, a rule-based automatic spell-checker was developed. Rules included the following: (1) If a character is used more than twice consecutively, then it is being emphasized in a manner that is likely not representative of actual pronunciation, therefore the editing process reduces any instance of three or more consecutive letters to two. For example, if a user comments "This video is cooool," the word "cooool" is replaced with "cool." (2) numbers are used to represent certain Arabic phonemes, e.g., the number 3 is commonly used to represent a pharyngeal glide. Users of the Arabic Chat Alphabet almost always write vowels (even in places where the corresponding vowels would not be written in Standard Arabic Orthography), therefore consonants should never be repeated except in cases of germination. Since the consonants represented by numerals in Arabic Chat Alphabet rarely germinate, the editing process reduces an instance of two or more numbers to one. (3) The editing process also gets rid of punctuation, and sets all the words to lower case, in order to avoid separating a word's term frequency. (4) The editing process also gets rid of numbers that are not part of an Arabic Chat Alphabet written word, in order to focus all of the experiments on the occurrences of spelled-out lexical items, at the expense of digit sequences. Two sets of experiments were conducted. In the first set of experiments, punctuation and digit sequence normalization (steps 3 and 4) were conducted, but not spelling normalization (steps 1 and 2). In the second experiment, all steps of text normalization (1 through 4) were applied. We compare the likelihoods of the two datasets to analyze the effects of the rule-based simple spell correction system.

LDA models were trained using the software provided by [6]. The number of topics was set to k=5. The number of topic mixtures was set to $\alpha=50/k=10$ as suggested by [7]. The same settings were used for every experimental trial; EM Convergence was set to be 1e-06.

Experiments were performed using three different types of data-sets. One type is a dataset with comments from only Lebanese Videos, another type is a dataset with comments from only Egyptian Videos, and the last type is a mixture of Lebanese and Egyptian data. By testing both intra-dialect and inter-dialect corpora, it is possible to test the hypothesis that dialect-dependent variation decreases the predictability of a corpus: if LDA models the inter-dialect dataset with lower log likelihood than either of the intra-

dialect datasets, then the difference between these log likelihoods can be interpreted as a measure of the entropy introduced by inter-dialect variability. Such a difference can be interpreted as support for the use of an Arabic Chat Alphabet as a bridge between the written forms of standard Arabic and the spoken forms of the Arabic Dialects, with the potential to, improve Arabic Speech-to-text Transcription.

4 Experimental results

LDA models were constructed in a three-fold cross-validation paradigm; log likelihood convergence plots from the three folds are shown in Figures 1, 2, and 3. In each fold of the cross-validation test, LDA models were trained for the inter-dialect dataset, the intra-dialect Egyptian Arabic dataset, and the intra-dialect Lebanese Arabic dataset. As the LDA algorithm converges toward a topic distribution in each ease its asymptotic log likelihood is consistently (across all three folds of cross-validation) higher for intra-dialect datasets than for the dataset whose text is a mixture of different dialects. This is consistent in all 3 tests. In all 3 tests, each dataset with the test (Lebanese, Egyptian, and Mixed) had almost the same numbers of unique terms. This was in order to eliminate variance caused by size rather than dialect. Our tests show that phonetic Latin spelled Arabic online chat is a dataset that encompasses different written languages from different dialects.

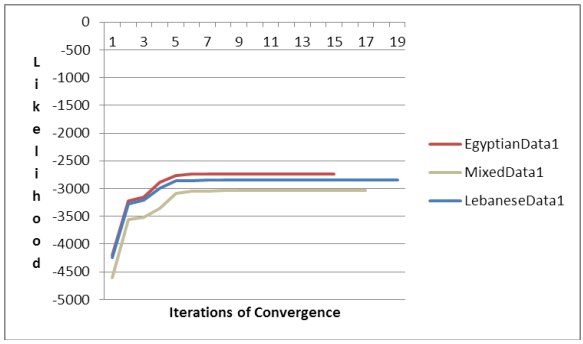

Figure 1: Log likelihood of the LDA model as a function of iteration number, plotted for each dialect and for the mixed-dialect data, fold 1

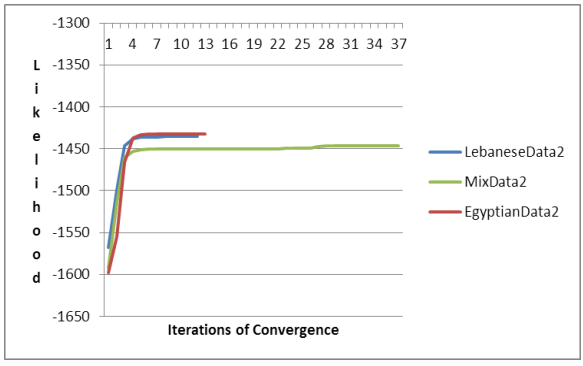

Figure 2: Log likelihood of the LDA model as a function of iteration number, plotted for each dialect and for the mixed-dialect data, fold 2

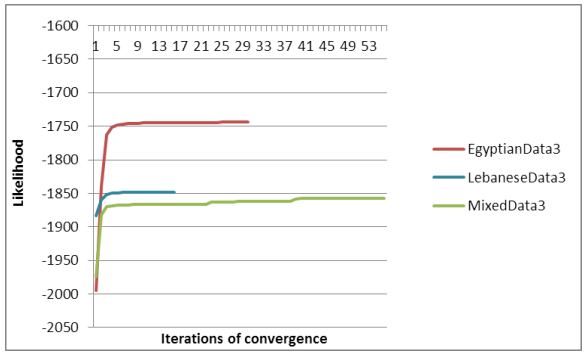

Figure 3: Log likelihood of the LDA model as a function of iteration number, plotted for each dialect and for the mixed-dialect data, fold 1

The second set of experiments evaluated the effect of a preliminary rule-based spelling normalization. LDA models were created for the inter-dialect corpus both with and without the normalization of repeated letters. Figure 4 shows the log likelihood convergence rates of LDA models trained with and without rule-based normalization of repeated spellings. As can be observed in figure 4, the spell checking system described in experimental methods did not increase the relation and topics modeling ability of the data.

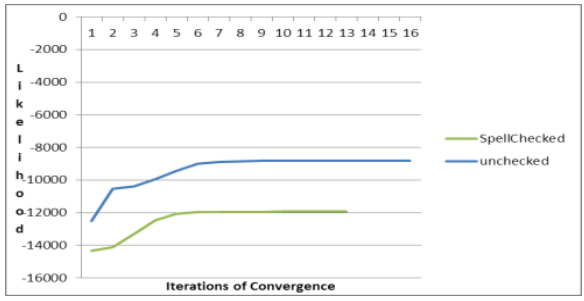

Figure 4: Log likelihood of the LDA model as a function of iteration number, spellchecked vs. unchecked data

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Beautiful(N)	I did(N)	Because (E)	Throne (N)	We're doing(N)
Enough(N)	Beautiful(N)	Charity(N)	We're doing (N)	Because (E)
Bored(N)	Kiss (N)	Truth(N)	Someone (E)	Beautiful(N)
Seriously(E)	Grow(N)	The Peace(N)	Seriously (E)	Better(N)
Blood(N)	Point to you(N)	Dreams(N)	Those (E)	Dreams(N)
Can you (L)	Blood(N)	I Dream(N)	Algerian (E)	Loved(N)
New(N)	Honey(N)	I love you(N)	Calm(N)	Sympathetic (E)
With her (L)	We will win (E)	Too Much(N)	Lebanese (L)	You(N)
Not(N)	With Regret(N)	Hide you(N)	With regret(N)	New(N)
See you(N)	Sundown (E)	Peace(N)	Not(N)	Wish(N)
What (L)	See(N)	Not(N)	Leave me (E)	Second(N)
Voice(N)	2 Egyptian	Traveled(N)	See you(N)	Peace(N)
Hey(N)	0 Lebanese	The Egyptians(E)	God keeps you from us(N)	God keeps you safe(L)
God Protects you(N)	9 Neutral	Day(N)	God Protects you(N)	God Protest you(N)
Nice(N)		Yours(N)	He curses her (E)	Stole it(N)
1 Egyptian		2 Egyptian	6 Egyptian	Sun(N)
3 Lebanese		0 Lebanese	1 Lebanese	Leaves(N)
11 Neutral		13 Neutral	8 Neutral	And the(N)
5 incomprehensible or not Arabic	9 incomprehensible or Not Arabic	5 incomprehensible or not Arabic	5 incomprehensible or not Arabic	2 Egyptian, 1 Lebanese, 17 Neutral ,1 incomprehensible

Figure 5: Most frequent intelligible word tokens in each of the five topics resulting from LDA analysis of the music discussion corpus.

Figure 5 lists the words of each of the 5 topics that are both comprehensible and also within the 20 most common terms (comprehensible or not) of the topic. The word lists seem to show some dialect-dependent tendencies. In Topic 3, all of the comprehensible words are entirely Egyptian or dialect-neutral. In Topic 1, 14/15 of the comprehensible words are neutral or Lebanese.

5 Discussion

Results of this experiment demonstrated that (1) dialect-dependent variation consistently reduces the log likelihood with which LDA is able to model a text corpus, (2) LDA topics computed from the inter-dialect corpus sometimes contain words from only one dialect, and sometimes contain words from both dialects.

A simple rule-based spelling normalizer was tested for these data, but did not increase log likelihood of the model. It is self-evident, on even a casual perusal of the data, that considerable spelling variation exists, but apparently the simple rule-set tested here was insufficient to improve the log likelihood of the LDA model. Future work should evaluate better spell checking methods.

One possible method for improved spelling normalization would be a translation library in which different spellings of a word are automatically identified, e.g., based on weighted Levenshtein distance among words with similar n-gram statistics. One could then utilize the translation library, through regular expression manipulation, to search a large corpus to find which spelling of each word is the most common. The most common could then be accepted as the most likely correct spelling. This solution would in most cases select the spelling that is most similar to the spoken version of that word in the target dialect.

Our tests of Egyptian and Lebanese Online chat data suggest that this form of data collection, collecting online chat data, could be very useful for transcribing Arabic Dialects. It has been argued elsewhere that improved Arabic dialect transcription could significantly improve Arabic Speech Recognition. This data collection could also be useful in internet search. Latin-spelled Arabic chat data are not currently utilized ideally in Arabic Internet search, because the data has no uniform spelling, and because the vast majority of Arabic text (on-line or elsewhere) is spelled in standard Arabic using the Arabic alphabet.

Another confounding variable to the data that we collected from online chat is the potential bias of topic. We collected YouTube comments from different videos' comment sections. All the videos were about music. Only comment sections on the topic of music were selected because of the fear that topic-dependent, rather than dialect-dependent variation could be the cause of the differences in Lebanese and Egyptian text. Despite our attempts, such variation is unlikely to be completely absent.

Future work, that could springboard off of the research reported in this paper, includes the testing of large corpora of parallel text of standard written Arabic, Egyptian speech-like text, and Lebanese speech-like text (the Egyptian and Lebanese speech-like text both being similar to the data collected for this paper). If one were to collect Egyptian online chat and then translate it into standard written Arabic, and do the same with Lebanese online chat, one would get four datasets. If it is assumed that all, or almost all, of the discussion dependent variation is contained within the variation between the standard written Arabic derived from Egyptian online chat, and the standard written Arabic derived from Lebanese online chat, then one could test for dialect-dependent variation in online chat (i.e. Arabic speech-like text) by the following. Measuring variation between the Lebanese and Egyptian online chats, and then seeing whether or not this variation is consistently greater then the variation between these same two sets of text translated into standard written Arabic.

Future work into developing a sufficient spell checking system for Latin Spelled Arabic, could allow Arabic online chat data to be utilized effectively in Arabic search queries. This may be true, not just of Arabic, but also of many other non-Latin based languages, like Mandarin and Hindu, that also have copious amounts of Latin Spelled chat data online. Our novel data collection source and the results of LDA analysis on our corpus show that Latin Spelled Phonetic Arabic from Online chat can be a very useful information source for Arabic Speech-to-text Transcription, Speech Recognition, and Information Retrieval.

6 Conclusion

In this paper we utilize LDA, a generative probabilistic modeling method, to analyze a phonetic Latin Spelled Arabic online chat corpus. Our experimental Results seem to indicate that phonetic Latin Spelled Arabic contains dialect dependent variation. This dialect dependent variation can potentially aid in producing written forms of Arabic Dialects, a contemporary difficulty in Arabic language processing.

7 Acknowledgements

This research was partially supported by grant number NPRP 410-1-069 from the Qatar National Research Fund.

8 References

[1] Lamel L., Messaoudi, A., and Gauvain, J-L. 2009. Automatic Speech to Text Transcription in Arabic.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of Machine Learning research 3 (2003): 993-1022.

[3] Jen-Tzung Chien and Chuang-Hua Chueh. Latent Dirichlet Language Model for Speech Recognition. 2008 IEEE Workshop on Spoken Language Technology.

[4] Ying-Lang Chang and Jen-Tzung Chien. Latent Dirichlet Learning for Document Summarization. 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[5] M. Elmahdy, R. Gruhn, S. Abdennadher and W. Minker. Rapid Phonetic Transcription using Everyday Life Natural Chat Alphabet Orthography for Dialectal Arabic Speech Recognition. 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[6] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Neural Information Processing Systems 16*, 2003.

[7] Griffiths and M. Steyvers. Finding Scientific Topics, Proc. of the National Academy of Sciences. PNAS April 6, 2004 vol. 101.

[8] Blei, David M., and John D. Lafferty. "Topic models." Text mining: classification, clustering, and applications 10 (2009): 71.

[9] Blei DM, Franks K, Jordan MI, Mian IS (2006) Statistical modeling of biomedical corpora: Mining the Caenorhabditis genetic center bibliography for genes related to life span. BMC Bioinformatics 7: 250.

[10] Kim, Samuel, Shrikanth Narayanan, and Shiva Sundaram. "Acoustic topic model for audio information retrieval." Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on. IEEE, 2009.

[11] Blake, Robert. "Computer mediated communication: A window on L2 Spanish interlanguage." Language Learning & Technology 4.1 (2000): 120-136.