A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic

Hitham M. Abo Bakr
Computer & System Dept.
Zagazig University
hithamab@yahoo.com

Khaled Shaalan
The Institute of Informatics
The British University in Dubai
khaled.shaalan@buid.ac.ae

Ibrahim Ziedan
Computer & System Dept.
Zagazig University
i.ziedan@yahoo.com

Abstract

Recently the rate of written colloquial text has increased dramatically. It is being used as a medium of expressing ideas especially across the WWW, usually in the form of blogs and partially colloquial articles. Most of these written colloquial has been in the Egyptian colloquial dialect, which is considered the most widely dialect understood and used throughout the Arab world. Modern Standard Arabic is the official Arabic language taught and understood all over the Arab world. Diacritics play a key role in disambiguating Arabic text. The reader is expected to infer or predict vowels from the context of the sentence. Inferring the full form of the Arabic word is also useful when developing Arabic natural language processing tools and applications. In this paper, we introduce a generic method for converting a written Egyptian colloquial sentence into its corresponding diacritized Modern Standard Arabic sentence which could easily be extended to be applied to other dialects of Arabic. In spite of the non-availability of linguistic Arabic resources for this task, we have developed techniques for lexical acquisition of colloquial words which are used for transferring written Egyptian Arabic into Modern Standard Arabic. We successfully used Support Vector Machine approach for the diacritization (aka vocalization or vowelling) of Arabic text.

1. Introduction

Arabic is a Semitic language spoken by over 250 million people throughout the Middle East and North Africa. It is one of the six official languages of the United Nations. Arabic is the language of Islam and its holly book- the Qur'aan. It is also the language in which some of the world's greatest works of literature, science, and history have been written [1].

Development of Natural Language Processing (NLP) tools for Arabic has been hindered by:

- The highly inflected nature of the language and its complex linguistic structure. That is why many NLP tools and applications follow statistical-based approaches. The main challenge to these approaches is the requirement of large amount of training data which have only been recently available.
- Non-availability of linguistic resources. Although few corpora are available for Arabic they are expensive and sometimes they do not fit well with many NLP tasks. Researchers of Arabic NLP have to develop their own linguistic acquisition module(s) in order to be able to approach their researches.
- The existence of a family of dialects such that speakers of some of these dialects are unable to understand speakers of other Arabic dialects. Using Modern Standard Arabic (MSA) as a hub language, into and out of which all processing is done, will make the transfer among these Arabic colloquial dialects straight away.
- The impact of non-Arabic words. As we know, most of colloquial words are derived from Arabic words[2]. One factor in the differentiation of the colloquial is influenced by languages previously spoken in the areas like Turkish language [3], which have typically provided a significant number of new words, and have sometimes also influenced pronunciation or word order.
- Lacks of diacritics in MSA text. Diacritics sometimes used to disambiguate words. The reader is expected to infer or predict vowels

INFOS2008, March 27-29, 2008 Cairo-Egypt

from the context of the sentence. Inferring the full form of the Arabic word is also useful when developing Arabic natural language processing tools and applications.

Nowadays, the rate of written colloquial text has increased dramatically. It is being used as a medium of expressing ideas especially across the WWW, usually in the form of blogs and partially colloquial articles. Most of these written colloquial has been in the Egyptian colloquial dialect, aka *Masri*, which is considered the most widely dialect understood and used throughout the Arab world. For this reason we selected Egyptian Colloquial Arabic to prove the capability of our approach in producing diacritized MSA from an input written Colloquial Arabic. MSA is the official Arabic language taught and understood all over the Arab world. It is generally written without diacritics. For NLP purposes, it better to get the output text represented in diacritized form in order to disambiguate words.

In this paper, we introduce a generic method for converting a written Egyptian colloquial sentence into its corresponding diacritized MSA sentence which could easily be extended to be applied to other dialects of Arabic. In spite of the non-availability of linguistic Arabic resources for this task, we have developed techniques for lexical acquisition of colloquial words which are used for transferring written Egyptian Arabic into Modern Standard Arabic. We have built a colloquial lexicon [4] on top of an existing MSA lexicon[4,5]. Entries of the colloquial lexicon were acquired using a rule-based approach from a large amount of data across the WWW. The lexicon entry has the following features: 1) it contains the colloquial word along its corresponding MSA word(s), and 2) the correct order of the MSA word in the target sentence if it is replaced by its input colloquial word. We used Support Vector Machine (SVM) approach [6,8] for the diacritization (aka vocalization or vowelling) of Arabic text. For example, the proposed system takes a colloquial sentence such as "لقيت الفلوس فين؟" (did-you-find the-money where?) and produces the output " أَيْنَ وَجَدَت النُقُود؟" (Where did you find the money?). SVM is a supervised learning algorithm that is mathematically proved and has led to high performances in many NLP tasks [7;9].

In a previous work [10], which is closely related to our work, SVM approach is used to automatically tokenize (segmenting off clitics), part-of speech (POS) tagging and annotating of base phrases (BPs) in MSA text. The training files were created from Arabic TreeBank[1] (ATB version 2.0) by extracting some features from the Treebank such as the POS for each segment of a sentence. This work was based on trained MSA corpora. In our work, we address the challenge of dealing with colloquial Arabic.

We adopted a tagging perspective for the three tasks. Thereby, we address them using the same SVM experimental setup which comprises a standard SVM as a multi-class classifier. The difference for the three tasks lies in the input, context, and features[7;10].

Due to the unavailability of Egyptian Colloquial Treebank (ECT), to the best of our knowledge, we had to build such an annotated data as it is needed for the learning purposes. In our approach we did not need a representation of a full tree as constructed in Arabic Penn Treebank. We just need the correct tokenization of each word and the POS for each segment used in training. So, we provide a demonstration of how to build the new annotated data. In our training, we distinguish between MSA words and Colloquial words. We used a package based on ATB for getting the correct tokenization (TOK) of the MSA word and the POS for each segment of a sentence. Colloquial words are handled in a semi-automatic way. Our main objective is to apply a hybrid approach for converting written Egyptian colloquial dialect into diacritized MSA. We apply a statistical-based approach for POS tagging (with enlarged tag set including MSA and Colloquial tags) and then apply rule-base approach for converting Egyptian Arabic words into their corresponding MSA words and diacrtizing these words. The advantage of the hybird approach is that we can distinguish between the correct Colloquial alternatives using the information provided by Colloquial POS tagger. As it is the case that colloquial word can take more than one possible meaning in MSA but with different POS. For example, the word "بقى" (Do something! or How about something?) can be used to indicate either an exclamation or an interrogative and could take the symbol "! or ?" at the end of the sentence. This is best explained by the following two examples: the input sentence "بقى أنت تعمل كدة؟" is to be transferred to the MSA as " أنت تفعل هذا!" (Do you do this!) but the input sentence "ازيك بقى؟" is to be transferred to the MSA as "كيف أخبارك؟" (How are you?).

[1] Treebank Corpus reference: http://www.ircs.upenn.edu/arabic/

We put an assumption that variants of colloquial words are mapped to a single canonical word form. For example, the words علشان ، علهشان ،عشان are all mapped to the word (علشان) that is stored in the Arabic lexicon. Another assumption is that input words are written correctly, i.e. free of spelling mistakes. The present system is designed to cover colloquial Egyptian words that would appear in written form. However, in some regions, such as Upper Egypt, they have different methods of pronunciations that are rarely used in written form and as such they are not covered by the current version.

The paper is structured as follows. Section 2 describes the steps followed in building the proposed system. Section 3 discusses the evaluation of the proposed system. In Section 4, we present some concluding remarks and give directions for future work.

2. Overview of the proposed System

In this section, we explain the steps for developing and evaluating the proposed system. They are:

1- Building colloquial lexicon and colloquial training corpus. In this step we collected 41705 words containing 9085 distinct non MSA words, 3000 distinct colloquial words, and the rest include spelling mistake words or non Arabic names. We kept only correct written sentences in order to be used by the training process.
2- Building training and test files. So, far, at the writing of this paper, we have created a test set with 800 sentences that ranges in size from 3 to 20 words. For the testing purposes another 200 sentences is to be used.
3- Conducting a training/test process. For this purpose we used yamCha-0.33 tool[2] for training and testing purposes.
4- Building diacritized MSA sentences. This is done by building a utility to convert a Colloquial word with the corresponding MSA word and takes care about the proper order of the target words in the produced MSA sentence.

2.1. Building colloquial lexicon and colloquial corpus

For details about building colloquial lexicon and colloquial training corpus we refer the reader to [4]. This step is briefly described as follows:

1- Create colloquial corpus, this done by downloading different pages from the WWW by using the freeware GNU Wget[3] for retrieving files using the most widely-used Internet protocols: HTTP, HTTPS and FTP.
2- Extract the Arabic sentences from downloaded pages by using text extraction utility developed for this purpose.
3- Recognize sentences that include colloquial words.
4- Add new colloquial words to the colloquial lexicon.

2.2 Building training files and colloquial corpus

In this step the training include both MSA words and Colloquial words which are handled differently as follows.

1- Segment words and annotate the segments with the correct POS for all training sentences. A statistical package [10] for MSA is used to get the correct TOK and POS for all MSA words. For colloquial words, it must be processed manually as there is no such automated tool available.
2- Verify the annotation manually to check up and correct any mistake in the training files.

From the collected annotated sentences, 80% are used for statistical training and the rest will be used for the test purposes.

2.3 Conducting a training process

In this step, we used the Colloquial Arabic lexicon that we built on top of the Buckwalter's morphological analysis tool. We used this morphological analyzer tool because the used Treebank data is built using this morphological analyzer.

1- The training is done using YamCha tool to create the Tokenization data used in TOK process task. This process is responsible for segmenting the word into clitics. The input is an Arabic windows code page (1256) text that is converted to Bulkwalter's transliteration code page. Finally, the text is converted to YamCha format [7]. Each word of the trained data is segmented to one of the following classes: Prefix1, prefix2 , Prefix3,word1,suffix1.
2- The POS process task is responsible for classifying clitics to one of 42 Classes of

[2] http://chasen.org/~taku/software/yamcha/

[3] http://www.gnu.org/software/wget/

Arabic POS[4]: CC, CD, DT, FW, IN, JJ, NN, NNP, NNPS, NNS, PRP, PRP$, PUNC, RB, RP, UH, VBD, VBN, VBP, WP, and WRB [11]. The colloquial tags for colloquial words are: Q_CC, Q_CD, Q_DT, Q_FW, Q_IN, Q_JJ, Q_NN, Q_NNP, Q_NNPS, Q_NNS, Q_PRP, Q_PRP$, Q_PUNC, Q_RB, Q_RP, Q_UH, Q_VBD, Q_VBN, Q_VBP, Q_WP, and Q_WRB.

2.4 Building diacritized MSA sentences

Running the system using 20% of the collected annotated sentences. This step is briefly described as follows:

1- Apply the tokenization task, see Figure 1. The system input is an MSA sentence containing some Colloquial words written with 1256 code page, The first step is to convert the 1256 code page to Buckwalter's transliteration (a Romanized representation). The Second step is to convert the transcribed sentence into YamCha format file to properly suit YamCha tool. Generally speaking, training and test file must consist of multiple tokens. The definition of tokens depends on the desired task.

In tokenization training files, the first column contains the typical letters of the word. The second column is true answer tag associated IOB2 model. In this model, three tags are used: namely, I, O and B to indicate inside, outside and beginning of a chunk, respectively. The output of YamCha training includes a new column at the end that describes the correct tokenization. In Figure 1, tokenization of the word ‘Alflws’ are two tokenized entities the prefix ‘Al’ followed by the word ‘flws’. The final output for this sentence consists of 5 tokenized entities: ‘lqyt’ , ‘Al’, ‘flws’, ‘fyn’ and ‘?’.

2- Apply the POS task, see Figure 2. The input for this step is the tokenized sentence; the first block contains two columns: a token with an initial dummy POS. The second block in Figure 2 contains the result after the testing using SVM, the third column shows the correct POS output produced by the YamCha.

After the system generates a POS for each token statistically by YamCha tool, the following rule-based steps will be applied to get the diacritized word from its POS. We distinguish between the handling of MSA words as opposed to Colloquial words:

- For each colloquial word, a lookup procedure is applied on the colloquial lexicon to get the corresponding diacritized MSA word and its correct position in the target sentence (e.g. the word فين 'fyn' will have the corresponding MSA word أَيْنَ' >ayona' (where) and its position should be at the start of the sentence).
- For each MSA word, a lexicon lookup is performed (by ignoring the results of the case ending) to get different alternatives for the same word along with its corresponding POS. These alternatives are compared with the results obtained from the POS learning. The correct diacritized word is chosen according to its POS. In some cases during the lexicon lookup we got more than one result for the same word and its POS that were produced from the training process. The input word "معلم" can have two diacritization alternatives with the same POS: "مُعَلّم" (a teacher - noun) or "مَعْلَم" (sign;mark - noun). This kind of ambiguity is resolved by getting the correct diacritization using a diacritized bigram database that we extracted from "ATB ver2" in addition to diacritized data acquired from some books published electronically.

[4] http://www.ircs.upenn.edu/arabic/Jan03release/arabic-POStags-collapse-to-PennPOStags.txt

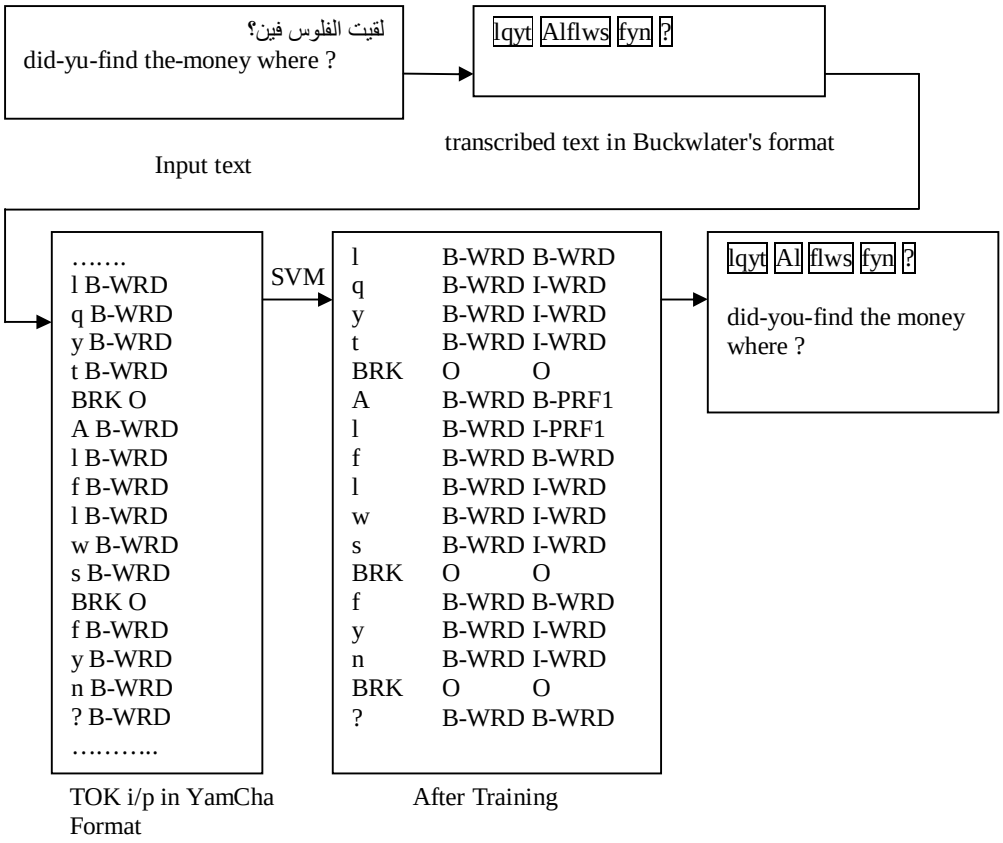

Figure 1: Tokenization process task

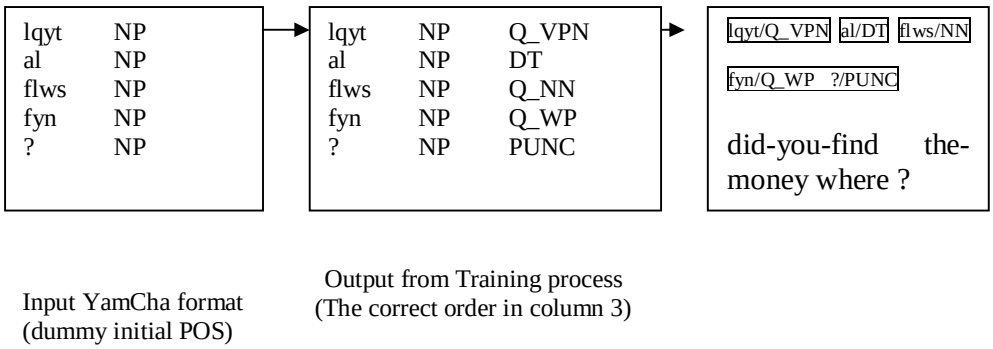

Figure 2 : POS process task

3- Evaluation

We conducted an evaluation using 1000 colloquial Egyptian sentences with 800 used for learning the system and 200 sentences for evaluating it. The objective is to test the accuracy of: 1) converting colloquial Arabic into MSA, 2) tokenization and POS tagging of both colloquial Arabic and MSA, and 3) diacritizing the final MSA output. We have no colloquial data available to compare with but there are some figures on the literature that has been used with the same training tool for TOK and POS but applied only MSA data. We will take these figures as an indicator of our success. It is worth mentioning that we have also a new POS tag sets for colloquial Arabic.

The evaluation sequence is as follows:

1- Apply the statistical method on the test set to get the TOK and POS annotations. These results are compared with manual annotation for the same set.
2- Apply the rule-based method for generating the diacritized MSA sentence. The accuracy is computed manually by three Arabic linguists to determine whether the correctness of: a) converting colloquial Egyptian sentence into MSA sentences, b) assigning the diacritic signs and c) producing the words into their proper order within the sentence.

The accuracy of converting Egyptian Colloquial Arabic words into their corresponding MSA words shows that 88% were correct. The accuracy of TOK and POS shows that 90% and 85% were correct, respectively. These results are comparable to those obtained from application on MSA data in [10]. The accuracy of assigning the diacritic signs shows that 70% were correct and the accuracy of producing the words into their proper order shows that 78% were correct.

4- Conclusions and Future work

We have presented a generic approach for converting an Egyptian colloquial Arabic sentence into a diacritized MSA sentence using a hybrid approach that combines a statistical approach that automatically tokenizes and tags Arabic sentence with a rule-based approach that constructs a target diacritized MSA sentence.

We have introduced new POS tags to annotate colloquial data. Adding more tags need to be investigated in order to improve the performance of the system. We can also achieve better results by increasing the size of the training corpus.

The presented approach is language independent and highly accurate. The results show that it is promising and could be used with other colloquial languages such as Syrian [1] or gulf colloquial[5]. We hope that this research could help colloquial Arabic speakers to convey their thoughts in a formal way which improves the human-to-human communication.

References

[1] Mary-Jane Liddicoat, Richard Lennane, and lman Abdul Rahim, Syrian Colloquial a functional course Arabic National Library of Australia Cataloguing-in-Publication entry, 1998.

[2] Shawki Deef, Tahrifat Al Amiah Lil Fousah Fi El Kawaad wa Al Bonian we Al Horouf wa Al Harakat , تحريفات العامية للفصحى في القواعد والبئيات والحروف والحركات , Dar El Maaref, Egypt, 1994.

[3] Abd El Sabor Shahin et al, Research and studies in Arabic dialects - Turkish uses in the Colloquial Egyptian, بحوث ودراسات في اللهجات العربيّة - استعمالات تركية في العامية المصرية, Academy of Arabic Language, Cairo, Egypt, http://www.arabicacademy.org.eg/FrontEnd/PrintDetails.aspx?PKPrintingTypeID=20 .

[4] Khaled Shaalan, Hitham M.Abo Bakr, and Ibrahim Ziedan, "Transferring Egyptian Colloquial Dialect into Modern Standard Arabic ", International Conference on Recent Advances in Natural Language Processing (RANLP – 2007) , Borovets, Bulgaria, PP. 525-529, September 27-29, 2007.

[5] Tim Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0. ", Tim Buckwalter ed Linguistic Data Consortium: University of Pennsylvania, 2002.

[6] Taku Kudo and Yuji Matsumoto, " Fast methods for kernel-based text analysis," In Proceedings of the 41st Annual Meeting on Association For Computational Linguistics - Volume 1 (Sapporo, Japan, July 07 - 12, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, 2003.

[5] http://www.alamuae.com/uaedic/pages-1.html

[7] Taku Kudoh and Yuji Matsumoto, "Use of Support Vector Learning for Chunck Identification," In Proceedings of the 4th Conference on CoNLL-2000 and LLL-2000, pages 142--144, 2000.

[8] Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods The Press Syndicate of the University of Cambridge, Cambridge, United Kingdom, 2000.

[9] Marti A. Hearst, "Support Vector Machines," IEEE Intelligent Systems, vol. 13, no. 4, pp. 18-28, Jul/Aug, 1998.

[10] Mona Diab, Kadri Hacioglu, and Daniel Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," In Proc. of HLT/NAACL 2004, Boston,2004.

[11] Beatrice Santorini, "Part-of-Speech Tagging Guidelines for the Penn Treebank Project," ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz, 1990.