Transferring Egyptian Colloquial Dialect into Modern Standard Arabic

Khaled Shaalan
The Institute of Informatics
The British University in Dubai,
PO Box 502216, Dubai, UAE
khaled.shaalan@buid.ac.ae

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

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

Abstract

Arabic is rooted in the Classical or Qur'anical Arabic, but over the centuries, the language has developed to what is now accepted as Modern Standard Arabic (MSA). Arab colloquial dialects are generally only spoken languages, but recently the rate of colloquial written text increases dramatically 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. We are able to reuse MSA processing tools with colloquial Arabic by transferring colloquial Arabic words into their corresponding MSA words. The advantages of this lexical transfer are to facilitate the communication with colloquial Arabic speakers and restoring it to the standard language in use nowadays. This paper addresses the transfer techniques between colloquial Arabic and MSA, which have not yet been closely studied before. In particular, we present a rule-based lexical transfer approach for converting Egyptian colloquial words into their corresponding MSA words. This process involves morphological analysis and lexical acquisition of colloquial words.

Keywords

Colloquial Arabic dialects processing, and transferring Egyptian Arabic into Modern Standard Arabic.

1. Introduction

Colloquial Arabic is a collective term for the spoken languages or dialects of people throughout the Arab world. Although it is descended from Arabic, it is considered a separate language. Speakers of some of these dialects are unable to understand speakers of other Arabic dialects. Recently, the rate of colloquial written text increases dramatically. Modern Standard Arabic (MSA) is the official Arabic language taught and understood all over the Arabic world. MSA has many challenges concerning the development of morphological and syntactic processing tools. These significant tools will become more complicated if they include in parallel the handling of Colloquial Arabic problems.

Today Egyptian Arabic, also known as *Masri*, is the dialect spoken in Egypt by more than 70 million people. It is understood across the Middle East due to the predominance of Egyptian media, making it one of the most widely spoken and most widely studied varieties of Arabic. For this reason we selected Egyptian Arabic to prove the capability of our approach in transferring a Colloquial Arabic dialect into MSA.

In literature, there are few researches that relate colloquial Arabic to MSA [6, 7]. These researches have focused on the spoken colloquial features of Arabic while our research focuses on written colloquial Arabic. Our approach is to develop transfer techniques that are able to perform the lexical mapping between written colloquial Arabic and MSA. The resultant front-end module will make it easy to incorporate colloquial Arabic into existing MSA tools. This will widen the coverage of current Arabic natural language processing applications to include colloquial languages or dialects of Arabic. Our proposed research builds the linguistic transformation resources between colloquial Arabic and MSA using the rule-based method. The data collection process will gather colloquial words from Arabic websites across the Web.

The paper is structured as follows. Section 2, discusses the challenges in handling written colloquial Arabic. In Section 3, we propose solutions for these problems. Section 4 gives background information. Section 5 concentrates on handling the deviation of Egyptian Arabic from MSA. Section 6 gives some concluding remarks.

2. Challenges in Handling Written Colloquial Arabic with Regard to MSA

Language processing of colloquial Arabic is a difficult task. The reasons of this difficulty come from several sources:

1- *Arabic Script*. There two ways that colloquial Arabic speaker use in their writing of colloquial words. One way is to Romanize the colloquial word (written using the Latin alphabet) and hence has to be transliterated from Arabic to English. Informal chatting across chat rooms or exchanged SMS messages in the Arab community usually done using Romanized letters. The other way is to write Arabic words using lexographic Arabic letters. Colloquial normal Arabic letters.

2-*Deviation from MSA*. There are five main deviations from MSA:

- Distortion of verbs (e.g. بليته من بللته – ضَرَبْتِيه من ضَرَبْتِه - حاكتب من سأكتب - ماتأعد من أما تقعد).
- Distortion of nouns. (e.g. الخِير من الخَير - ده من هذا - جَمهور خايف من خائف - من جُمهور- مين من مَنْ - فين من أين).
- Distortion of Pronouns and letters meanings. (e.g. عصايتي من عصاي - احنا من نحن - هوَّ من هُوَ).
- Distortion of the structure of the word form (e.g. اتاوب من تثاءب - اتَّاوى من اوى - بغبغان من ببغاء - تلات شهور من ثلاثة شهور).
- Replace the characters and movements. (e.g. تِعبان من ثَعبان - توم من ثوم - سقب من ثقب - شبط من شبث "أي تعلق").

3- *Lack of syntactic rules*. There are no identified grammar rules for colloquial dialects.

4-*Lexical expansion rate*. As colloquial Arabic is more popular than MSA, it is very often to observe much more newly added expressions/words as apposed to MSA.

3. The Proposed Approach

For the problems introduced in the previous section, we give suggestions for each of which.

To solve problem of writing colloquial Arabic in Latin alphabet, we propose the following process:

- Detect Romanized words in the input and transliterate theses words into Arabic lexographic letters,
- Normalize the words such as removing repeated characters that is usually used to informally indicate emotions, and
- Lookup the Colloquial-to-MSA lexicon for the closest colloquial word match and return the corresponding colloquial entry.

As an example, the phrase “Meeeesh 3aweez 7agh” will be converted to “ميش عاوز حاجة” (I do not need anything).

To solve the problem of the deviation of Egyptian Arabic from MSA, the major contribution of this research, we used an existing mature MSA lexicon (Buckwalter lexicon version2, [3][1]) to build the Colloquial-to-MSA lexicon such that both their entries coexist in one lexicon. We followed the same morphological analysis approach of this tool in analyzing the colloquial Arabic word. A rule-based lexical transfer approach is use to transform the analyzed colloquial Arabic word into MSA word(s).

To solve the problem of the lack of identified colloquial syntactic rules, we suggest solving this problem with empirical corpus-based techniques from Example Based Machine Translation (EBMT) [8, 9]. This has incurred building a parallel corpus of both the colloquial and MSA text. The development of such corpus is relatively new and will be published elsewhere.

To solve the problem of acquiring new colloquial words/expressions, we propose a process based on EBMT techniques that maintains the lexicon and keeps it up-to-date. This sophisticated process will gather Arabic text from the Web. The text is analyzed in order to recognize the unknown lexical items. An Arabic specialist has to take a decision of whether or not to add the unknown lexical item to the lexicon.

4. The Buckwalter Morphological Analyzer

We build our system on top of Buckwalter Arabic Morphological Analyzer Version 2.0 [3]. His morphological analysis depends on a dictionary of prefixes, a dictionary of suffixes, a stem dictionary, and three checking tables for testing the validity of a word analysis. The

[1] See the description of the Buckwalter's Arabic morphological analyzer http://www.qamus.org/morphology.htm

morphological analyzer tries to breakdown the input Arabic word into three elements: prefix, stem, and suffix. If all the three word elements are found in their respective lexicons, then their respective morphological categories are used to determine whether they are compatible. If all the morphological category pairs are compatible, then the morphological analysis is valid.

Each entry in the three lexicon files consists of four tab-delimited fields:

1. the entry (prefix, stem, or suffix) without short vowels and diacritics,
2. the entry (prefix, stem, or suffix) with short vowels and diacritics,
3. its morphological category (used for the compatibility between prefixes, stems, and suffixes), and
4. its English gloss(es), including selective POS data within XML tags `<pos>...</pos>`

Only fields 1 and 3 are required for morphological analysis. Fields 2 and 4 provide additional information once the morphology analysis is succeeded in producing the analyzed word(s). Arabic script data in the lexicons is provided in the Buckwalter transliteration scheme.

The following is a description of the three lexicon files:

- *dictPrefixes* contains all Arabic prefixes and their concatenations. Sample entry:
  `w wa Pref-Wa <pos>wa/CONJ</pos>`
- *dictSuffixes* contains all Arabic suffixes and their concatenations. Sample entry:
  `p ap NSuff-ap [fem.sg] <pos>ap/NSUFF_FEM_SG</pos>`
- *dictStems* contains all Arabic stems. Sample entries:
  `ktb katab PV write`
  `ktb kotub IV write`

There are three compatibility tables; each of the three compatibility tables lists pairs of compatible morphological categories:

- Compatibility table *tableAB* lists compatible Prefix and Stem morphological categories, such as:
  `NPref-Al N`
  `NPref-Al N-ap`
- Compatibility table *tableAC* lists compatible Prefix and Suffix morphological categories, such as:
  `NPref-Al Suff-0`
  `NPref-Al NSuff-u`
- Compatibility table *tableBC* lists compatible Stem and Suffix morphological categories, such as:
  `PV PVSuff-a`

5. The Proposed Solution of Transferring Colloquial Arabic Dialect to MSA

Our proposed transfer techniques are based on previous studies of the transformations between the MSA and colloquial Arabic [1, 2, 4, 5]. We used the indicated variations to acquire the lexical transfer rules that can be used to derive the MSA word from a corresponding colloquial Arabic word. Additional rules will be acquired and judged by an Arabic specialist during the lexical acquisition process. These rules are used to analyze the input colloquial word and produce the target MSA word(s).

5.1 Examples of Egyptian Colloquial Word to MSA Transformations

The colloquial Arabic word is normally derived from a well-formed MSA word. This process can be traced back to the distortion (transformation) made to the MSA word that has changed it to a colloquial Arabic word form. The analysis of the relationship between well-formed MSA Arabic words and colloquial words has been discussed by many linguists [1, 2, 4, 5]. Table 1 shows distortion examples and how to transfer them into MSA words.

The transfer between Egyptian Arabic dialect and MSA is one-to-many transformation. This means some Egyptian Arabic words can be transferred in one or more steps through lexicon lookup as the mapping involves more than one morpheme. For example, the Egyptian word ازيك "How are you?" is transformed to two MSA words "كيف حالك؟". Other examples are:

- ماورد (Ma2 ward) : ماء ورد
- كلشينكان (Koleshenkan) : كل شيء كان
- أجرنك (2agranak): لا جرم انك وتقال في العامية أجرنك شاطر أي لا جرم انك شاطر
- أشمعنا (2eshMe3na): ايش المعني
- إكمنه (2kmeno): كما انه
- بسملة (Besmellah): بسم الله

In colloquial language processing, a word might be added to the lexicon which does not have a

corresponding word in the formal language. This is also the case in Egyptian colloquial (e.g. the word "بقى" can is used to indicate either an exclamation or an interrogation such that both the symbols "?!" appear together at the end of the sentence. This is best explained by the following examples:

- "بقى أنت تعمل كدة؟" which is transferred to MSA as "أنت تفعل هذا؟!" (Do you do this?), and
- "ازيك بقى؟" which is transferred to MSA as "كيف حالك؟!" (How are you?).

Table 1. Examples that illustrate the relation between MSA words and Egyptian Arabic words

MSA	EGW	Distortion Type	Handling method
يد	إيد	Replace of vowels " فتح الاول والعامية تكسره"	Add new stem and assign the same rules as (colloquial Arabic word (CAW)
وأنا نحن	ونا احنا	Distortion in Pronouns and letters meaning " التحريف في الضمائر"	Add new stem and assign the same rules as CAW
البارحة أمس	امبارح	Distortion in Pronouns and letters meaning " التحريف في حروف المعاني ال - ام"	Add new stem and assign the same rules as CAW "امس"to be suitable in MSA even it is more suitable "البارحة"
قال يا ليت متاع تَمَطّى سلحفاء	أال ياريت بتاع تمطع زحلفة	Replace the characters and vowels. " ابدال الحروف "	Add new stem and assign the same rules as CAW
ابتل ارتمى ارتوى اشتوى افتضح	اتبل اترمى اتروى اتشوى اتفضح	Distortions in the structure of the word " تقدم التاء علي فاء الفعل في صيغة أفتعل"	Add new stem and assign the same rules as CAW

5.2 Lexicon Structure

We enhanced the Buckwalter's lexicon tables with new extra fields:

- *ID*: An identifier to distinguish each word segment. This field is used for indexing purposes,
- *SegmentType*: it can be either MSA (Ar-Ar), Egyptian dialect (*Ar-Eg*) or other dialects such as Jordanian dialect (Ar-Jr) for future extension of the lexicon.
- *NewSegmentPosition*: this is the new position of the word segment, which indicates its proper order, within the target MSA word or sentence. This field takes one of the following values:
  - o same position (SP),
  - o start of word (SoW),
  - o end of word (EoW),
  - o start of sentence (SoS),
  - o end of sentence (EoS), and the like.

For example, the Egyptian colloquial sentence "جيت امتى؟" (you came when?) is literally transformed to the MSA sentence "جئت متى؟" (you came when?). Given that the word "امتى" takes the value "SoS" for the *NewSegmentPosition* field, the transformation moves this word to the beginning of the sentence in order to get the target MSA sentence " متى جئت؟" (When did you come?).

5.3 Mapping Rules

A new database file, called *Mapping Table (MT)*, is introduced to encode the mapping rules between Egyptian Arabic to MSA. This table uses the value of the lexicon's ID field to cross reference the lexical entries inside the rules. The mapping is either one-to-one or one-to-many. An entry of this table has three fields: source colloquial word, target colloquial word, and the mapping mode. The mapping mode takes either of two values: 0 indicates one-to-one and 1 indicates one-to-many. In the following we will present examples of mapping rules along with their related lexicon entries.

<u>Example 1:</u> mapping the colloquial interrogative "امتي" (when) to the MSA word "متي".

This rule will be represented in the MT by an entry with the values: source colloquial interrogative=ID 79831, target MSA interrogative ID=64063, and mapping mode=0, where the source and target words entries in the lexicon are:

```
64063   mtY     mataY   FW-Wa   when    مـــتى
        مَتَى   mataY/INTERROG_PART
        استفهام مَتَى/أداة      mataY_2
        2_مَتَى  mty(1) (متي1) SP Ar-Ar

79831   >mtY    >mtY    FW-Wa   when    أمتى
        أمَتَى  >mtY/INTERROG_PART
        استفهام أمَتَى/أداة     >mataY
        أمَتَى  >mty    أمتي    SoS  Ar-Eg
```

Example 2: mapping the colloquial prefix "عال" (on-the) to the MSA words "على" (on) and the prefix article "ال" (the).

These rules will be represented in the MT by two entries (one-to-many): 1) source colloquial prefix=ID 79835, target MSA preposition=46196, and mapping mode=0, and 2) source colloquial prefix=ID 79835, target MSA article=15, and mapping mode=1.

In addition to adding colloquial prefixes, stems and suffixes to the corresponding lexicon database file, the compatibility database files should also be modified to include entries that will verify the recognized prefix, stem, and suffix of the input Egyptian Arabic word. Consequently, the colloquial prefix "عال" (EAl) will have also entries in the respective compatibility tables: *tableAB*, and *tableAC*. As a matter of fact, these entries will be treated in a similar way to the MSA prefix "بال" (BiAl). In order to distinguish between MSA and colloquial entries, we used the prefix "C_" as an indicator of a colloquial entry, e.g. the morphological category of "عـال" (EAl) is "C_NPref-EAl" while the MSA for "بـال" (BiAl) is "NPref-BiAl"

6. Conclusion

We have investigated the variations between Egyptian Arabic and MSA, and introduced lexical transfer techniques between these languages. These techniques reuse existing Arabic morphological analysis resources and enhance these resources with meta data of Egyptian Arabic. Our approach is able to transfer written Egyptian colloquial dialect into its corresponding MSA forms in order to cope with the dramatic increase of written colloquial dialects. This step showed that it is easy to incorporate colloquial Arabic dialects into existing MSA tools. We hope these techniques to be applied to other colloquial Arabic dialects such as Moroccan, Levantine and Gulf Arabic. Moreover, using MSA Arabic as a hub language, into and out of which all transfer is done, will make the transfer among these Arabic colloquial dialects straight way such that speakers of one dialect is able to read and understand written material of other Arabic dialects.

References

1. Shawki Deef, Tahrifat Al Amiah Lil Fousah Fi El Kawaad wa Al Bonian we Al Horouf wa Al Harakat , تحريفات العامية للفصحى في القواعد والبنيات والحروف والحركات, Dar El Maaref, Egypt, 1994.
2. Scocrates Spiro ,"An Arabic – English Dictionary of the Colloquial Arabic of Egypt", Lebanon Bookshop Publisher, Lebanon, 1973
3. Tim Buckwalter, Buckwalter Arabic Morphological Analyzer Version 2.0 LDC Linguistic Data Consortium, University of Pennsylvania, 2004. Available at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004L02
4. Ahmed Taymour, "Moaagam Taymour Al Kbir: volume 1, 2 & 3", " معجم تيمور الكبير: مجلد 1 ؛ 2 ؛ 3", Dar El Afak el Arabia, Egypt, 2003.
5. Ibn El hanbaly, "Bahr ul-awwam fi ma asaba fihil a'wam, " بحر العوّام فيما أصاب فيه العوامّ", Ibn Zietoun, Syria,1937
6. Owen Rambow, David Chiang, Mona Diab and Nizar Habash, The final report: Parsing Arabic Dialects (version I), CSLP, JHU, Baltimore, USA, 2006.
7. Nizar Habash and Owen Rambow, MAGEAD:A Morphological Analyzer and Generator for the Arabic Dialects, In the Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, PP 681–688, July 2006.
8. Ralf D Brown, Example Based Machine Translation in the Pangloss System, In the proceedings of The 16th International Conference on Computational Linguistics, Copenhagen (COLING-96), pp 169-174, 1996.
9. Ralf D Brown and Robert Frederking Applying Statistical English Language Modeling to Symbolic Machine Translation, In the Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), Leuven, Belgium, pp 221-239, 1995.