Transforming Standard Arabic to Colloquial Arabic

Emad Mohamed, Behrang Mohit and Kemal Oflazer

Carnegie Mellon University - Qatar
Doha, Qatar
emohamed@qatar.cmu.edu, behrang@cmu.edu, ko@cs.cmu.edu

Abstract

We present a method for generating Colloquial Egyptian Arabic (CEA) from morphologically disambiguated Modern Standard Arabic (MSA). When used in POS tagging, this process improves the accuracy from 73.24% to 86.84% on unseen CEA text, and reduces the percentage of out-of-vocabulary words from 28.98% to 16.66%. The process holds promise for any NLP task targeting the dialectal varieties of Arabic; e.g., this approach may provide a cheap way to leverage MSA data and morphological resources to create resources for colloquial Arabic to English machine translation. It can also considerably speed up the annotation of Arabic dialects.

1. Introduction

Most of the research on Arabic is focused on Modern Standard Arabic. Dialectal varieties have not received much attention due to the lack of dialectal tools and annotated texts (Duh and Kirchoff, 2005). In this paper, we present a rule-based method to generate Colloquial Egyptian Arabic (CEA) from Modern Standard Arabic (MSA), relying on segment-based part-of-speech tags. The transformation process relies on the observation that dialectal varieties of Arabic differ mainly in the use of affixes and function words while the word stem mostly remains unchanged. For example**,** given the Buckwalter-encoded MSA sentence "*AlAxwAn Almslm**wn lm yfwzwA fy** AlAntxbAt*" the rules produce "*AlAxwAn Almslm**yn mfAzw$ f** AlAntxAbAt*" (الاخوان المسلمين مفازوش ف الانتخابات, The Muslim Brotherhood did not win the elections). The availability of segment-based part-of-speech tags is essential since many of the affixes in MSA are ambiguous. For example, ***lm*** could be either a negative particle or a question work, and the word ***AlAxwAn*** could be either made of two segments (***Al+<xwAn***, the brothers), or three segments (***Al+>xw+An***, the two brothers).

We first introduce the transformation rules, and show that in many cases it is feasible to transform MSA to CEA, although there are cases that require much more than POS tags. We then provide a typical case in which we utilize the transformed text of the Arabic Treebank (Bies and Maamouri, 2003) to build a part-of-speech tagger for CEA. The tagger improves the accuracy of POS tagging on authentic Egyptian Arabic by 13% absolute (from 73.24% to 86.84%) and reduces the percentage of out-of-vocabulary words from 28.98% to 16.66%.

2. MSA to CEA Conversion Rules

Table 1 shows a sentence in MSA and its CEA counterpart. Both can be translated into: *"We did not write it for them.*" MSA has three words while CEA is more synthetic as the preposition and the negative particle turn into clitics. Table 1 illustrates the end product of one of the Imperfect transformation rules, namely the case where the Imperfect Verb is preceded by the negative particle *lm*.

	Arabic	**Buckwalter**
MSA	لم نكتبها لهن	lm nktbhA lhn
CEA	مكتبنهلهمش	mktbnhlhm$
English	We did not write it for them	

Table 1: a sentence in MSA and CEA

Our 103 rules cover nominals (number and case affixes), verbs (tense, number, gender, and modality), pronouns (number and gender), and demonstrative pronouns (number and gender).

The rules also cover certain lexical items as 400 words in MSA have been converted to their com-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 176–180,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

mon CEA counterparts. Examples of lexical conversions include ***ZlAm*** and ***Dlmp*** (darkness), ***rjl*** and ***rAjl*** (man), ***rjAl*** and ***rjAlp*** (men), and ***kvyr*** and ***ktyr*** (many), where the first word is the MSA version and the second is the CEA version.

Many of the lexical mappings are ambiguous. For example, the word ***rjl*** can either mean ***man*** or ***leg***. When it means *man*, the CEA form is ***rAjl***, but the word for *leg* is the same in both MSA and CEA. While they have different vowel patterns (***rajul*** and ***rijol*** respectively), the vowel information is harder to get correctly than POS tags. The problem may arise especially when dealing with raw data for which we need to provide POS tags (and vowels) so we may be able to convert it to the colloquial form. Below, we provide two sample rules:

The imperfect verb is used, inter alia, to express the negated past, for which CEA uses the perfect verb. What makes things more complicated is that CEA treats negative particles and prepositional phrases as clitics. An example of this is the word ***mktbthlhm$*** (I did not write it for them) in Table 1 above. It is made of the negative particle ***m***, the stem ***ktb*** (to write), the object pronoun ***h*, the** preposition ***l***, the pronoun ***hm*** (them) and the negative particle ***$.*** Figure 1, and the following steps show the conversions of ***lm nktbhA lhm*** to ***mktbnhAlhm$:***

1. Replace the negative word ***lm*** with one of the prefixes ***m***, ***mA*** or the word ***mA.***
2. Replace the Imperfect Verb prefix with its Perfect Verb suffix counterpart. For example, the IV first person singular subject prefix **>** turns into ***t*** in the PV.
3. If the verb is followed by a prepositional phrase headed by the preposition ***l*** that contains a pronominal object, convert the preposition to a prepositional clitic.
4. Transform the dual to plural and the plural feminine to plural masculine.
5. Add the negative suffix ***$*** (or the variant ***$y***, which is less probable)

As alluded to in 1) above, given that colloquial orthography is not standardized, many affixes and clitics can be written in different ways. For example, the word ***mktbnhlhm$,*** can be written in 24 ways. All these forms are legal and possible, as attested by their existence in a CEA corpus (the Arabic Online Commentary Dataset v1.1), which we also use for building a language model later.

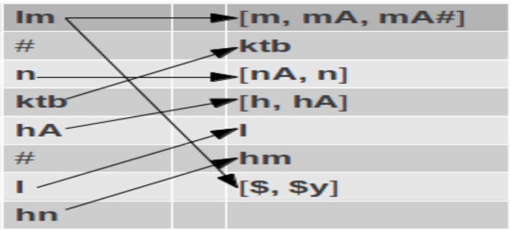

Figure 1: One negated IV form in MSA can generate 24 (3x2x2x2) possible forms in CEA

MSA possessive pronouns inflect for gender, number (singular, dual, and plural), and person. In CEA, there is no distinction between the dual and the plural, and a single pronoun is used for the plural feminine and masculine. The three MSA forms *ktAb**hm***, *ktAb**hmA*** and *ktAb**hn*** (***their** book* for the masculine plural, the dual, and the feminine plural respectively) all collapse to *ktAb**hm***.

Table 2 has examples of some other rules we have applied. We note that the stem, in bold, hardly changes, and that the changes mainly affect function segments. The last example is a lexical rule in which the stem has to change.

Rule	**MSA**	**CEA**
Future	swf y**ktb**	Hy**ktb**/hy**ktb**
Future_NEG	ln >**ktb**	m$ h**ktb/** m$ H**ktb**
IV	y**ktb**wn	by**ktb**w/ b**ktb**w/ b**ktb**wA
Passive	**ktb**	An**ktb/** At**ktb**
NEG_PREP	lys **mn**hn	m**mn**hm$
Lexical	**trk**hmA	**sAb**hm

Table 2: Examples of Conversion Rules.

3. POS Tagging Egyptian Arabic

We use the conversion above to build a POS tagger for Egyptian Arabic. We follow Mohamed and Kuebler (2010) in using whole word tagging, i.e., without any word segmentation. We use the Columbia Arabic Treebank 6-tag tag set: PRT (Particle), NOM (Nouns, Adjectives, and Adverbs), PROP (Proper Nouns), VRB (Verb), VRB-pass (Passive Verb), and PNX (Punctuation) (Habash and Roth, 2009). For example, the word ***wHnktblhm*** (*and we will write to them,* وحنكتبلهم*)* receives the tag ***PRT+PRT+VRB+PRT+NOM.*** This results in 58 composite tags, 9 of which occur 5 times or less in the converted ECA training set.

We converted two sections of the Arabic Treebank (ATB): p2v3 and p3v2. For all the POS tagging experiments, we use the memory-based POS tagger (MBT) (Daelemans *et al.*, 1996) The best results, tuned on a dev set, were obtained, in non-exhaustive search, with the Modified Value Difference Metric as a distance metric and with *k* (the number of nearest neighbors) = 25. For known words, we use the IGTree algorithm and 2 words to the left, their POS tags, the focus word and its list of possible tags, 1 right context word and its list of possible tags as features. For unknown words, we use the IB1 algorithm and the word itself, its first 5 and last 3 characters, 1 left context word and its POS tag, and 1 right context word and its list of possible tags as features.

3.1. Development and Test Data

As a development set, we use 100 user-contributed comments (2757 words) from the website ***masrawy.com***, which were judged to be highly colloquial. The test set contains 192 comments (7092 words) from the same website with the same criterion. The development and test sets were hand-annotated with composite tags as illustrated above by two native Arabic-speaking students.

The test and development sets contained spelling errors (mostly run-on words). The most common of these is the vocative particle ***yA***, which is usually attached to following word (e.g. ***yArAjl***, (you man, ياراجل)). It is not clear whether it should be treated as a proclitic, since it also occurs as a separate word, which is the standard way of writing. The same holds true for the variation between the letters ***** and ***z***, (ذ and ز in Arabic) which are pronounced exactly the same way in CEA to the extent that the substitution may not be considered a spelling error.

3.2. Experiments and Results

We ran five experiments to test the effect of MSA to CEA conversion on POS tagging: (a) **Standard**, where we train the tagger on the ATB MSA data, (b) **3-gram LM**, where for each MSA sentence we generate all transformed sentences (see Section 2.1 and Figure 1) and pick the most probable sentence according to a trigram language model built from an 11.5 million words of user contributed comments.[1] This corpus is highly dialectal Egyptian Arabic, but like all similar collections, it is diglossic and demonstrates a high degree of code-switching between MSA and CEA. We use the SRILM toolkit (Stolcke, 2002) for language modeling and sentence scoring, (c) **Random**, where we choose a random sentence from all the correct sentences generated for each MSA sentence, (d) **Hybrid**, where we combine the data in a) with the best settings (as measured on the dev set) using the converted colloquial data (namely experiment c). Hybridization is necessary since most Arabic data in blogs and comments are a mix of MSA and CEA, and (e) **Hybrid + dev,** where we enrich the Hybrid training set with the dev data.

We use the following metrics for evaluation: **KWA**: Known Word Accuracy (%), **UWA**: Unknown Word Accuracy (%), **TA**: Total Accuracy (%), and **UW**: unknown words (%) in the respective set in the respective experiment. Table 3(a) presents the results on the development set while Table 3(b) the results on the test set.

Experiment	**KWA**	**UWA**	**TA**	**UW**
(a) Standard	92.75	39.68	75.77	31.99
(b) 3-gram LM	89.12	43.46	76.21	28.29
(c) Random	92.36	43.51	79.25	26.84
(d) Hybrid	**94.13**	**52.22**	**84.87**	**22.09**

Table 3(a): POS results on the development set.

We notice that randomly selecting a sentence from the correct generated sentences yields better results than choosing the most probable sentence according to a language model. The reason for this may be that randomization guarantees more coverage of the various forms. We have found that the vocabulary size (the number of unique word types) for the training set generated for the **Random** experiment is considerably larger than the vocabulary size for the 3-gram LM experiment (55367 unique word types in **Random** versus 51306 in **3-gram LM**), which results in a drop of 4.6% absolute in the percentage of unknown words: 27.31% versus 22.30%). This drop in the percentage of unknown words may indicate that generating all possible variations of CEA may be more useful than using a language model in general. Even in a CEA corpus of 35 million words, one third of the words generated by the rules are not in the corpus, while many

[1] Available from http://www.cs.jhu.edu/~ozaidan/AOC

of these are in both the test set and the development set.

Experiment	KWA	UWA	TA	UW
(a) Standard	89.03	40.67	73.24	28.98
(b) 3-gram LM	84.33	47.70	74.32	27.31
(c) Random	90.24	48.90	79.67	22.70
(d) Hybrid	92.22	53.92	83.81	19.45
(e) Hybrid+dev	**94.87**	**56.46**	**86.84**	**16.66**

Table 3(b): POS results on the test set

We also notice that the conversion alone improves tagging accuracy from 75.77% to 79.25% on the development set, and from 73.24% to 79.67% on the test set. Combining the original MSA and the best scoring converted data (Random) raises the accuracies to 84.87% and 83.81% respectively. The percentage of unknown words drops from 29.98% to 19.45% in the test set when we used the hybrid data. The fact that the percentage of unknown words drops further to 16.66% in the **Hybrid+dev** experiment points out the authentic colloquial data contains elements that have not been captured using conversion alone.

4. Related Work

To the best of our knowledge, ours is the first work that generates CEA automatically from morphologically disambiguated MSA, but Habash et al. (2005) discussed root and pattern morphological analysis and generation of Arabic dialects within the MAGED morphological analyzer. MAGED incorporates the morphology, phonology, and orthography of several Arabic dialects. Diab *et al*. (2010) worked on the annotation of dialectal Arabic through the COLABA project, and they used the (manually) annotated resources to facilitate the incorporation of the dialects in Arabic information retrieval.

Duh and Kirchhoff (2005) successfully designed a POS tagger for CEA that used an MSA morphological analyzer and information gleaned from the intersection of several Arabic dialects. This is different from our approach for which POS tagging is only an application. Our focus is to use any existing MSA data to generate colloquial Arabic resources that can be used in virtually any NLP task.

At a higher level, our work resembles that of Kundu and Roth (2011), in which they chose to adapt the text rather than the model. While they adapted the test set, we do so at the training set level.

5. Conclusions and Future Work

We have a presented a method to convert Modern Standard Arabic to Egyptian Colloquial Arabic with an example application to the POS tagging task. This approach may provide a cheap way to leverage MSA data and morphological resources to create resources for colloquial Arabic to English machine translation, for example.

While the rules of conversion were mainly morphological in nature, they have proved useful in handling colloquial data. However, morphology alone is not enough for handling key points of difference between CEA and MSA. While CEA is mainly an SVO language, MSA is mainly VSO, and while demonstratives are pre-nominal in MSA, they are post-nominal in CEA. These phenomena can be handled only through syntactic conversion. We expect that converting a dependency-based treebank to CEA can account for many of the phenomena part-of-speech tags alone cannot handle

We are planning to extend the rules to other linguistic phenomena and dialects, with possible applications to various NLP tasks for which MSA annotated data exist. When no gold standard segment-based POS tags are available, tools that produce segment-based annotation can be used, e.g. segment-based POS tagging (Mohamed and Kuebler, 2010) or MADA (Habash et al, 2009), although these are not expected to yield the same results as gold standard part-of-speech tags.

Acknowledgements

This publication was made possible by a NPRP grant (NPRP 09-1140-1-177) from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

We thank the two native speaker annotators and the anonymous reviewers for their instructive and enriching feedback.

References

Bies, Ann and Maamouri, Mohamed (2003). Penn Arabic Treebank guidelines. Technical report, LDC, University of Pennsylvania.

Buckwalter, T. (2002). Arabic Morphological Analyzer (AraMorph). Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257- 0

Daelemans, Walter and van den Bosch, Antal (2005). Memory Based Language Processing. Cambridge University Press.

Daelemans, Walter; Zavrel, Jakub; Berck, Peter, and Steven Gillis (1996). MBT: A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, Proceedings of the 4th Workshop on Very Large Corpora, pages 14–27, Copenhagen, Denmark.

Diab, Mona; Habash, Nizar; Rambow, Owen; Altantawy, Mohamed, and Benajiba, Yassine. COLABA: Arabic Dialect Annotation and Processing. LREC 2010.

Duh, K. and Kirchhoff, K. (2005). POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, June 2005.

Habash, Nizar; Rambow, Own and Kiraz, George (2005). Morphological analysis and generation for Arabic dialects. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 17–24, Ann Arbor, June 2005

Habash, Nizar and Roth, Ryan. CATiB: The Columbia Arabic Treebank. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 221–224, Singapore, 4 August 2009. c 2009 ACL and AFNLP

Habash, Nizar, Owen Rambow and Ryan Roth. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009

Kundu, Gourab abd Roth, Don (2011). Adapting Text instead of the Model: An Open Domain Approach. Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 229–237,Portland, Oregon, USA, 23–24 June 2011

Mohamed, Emad. and Kuebler, Sandra (2010). Is Arabic Part of Speech Tagging Feasible Without Word Segmentation? Proceedings of HLT-NAACL 2010, Los Angeles, CA.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In Proc. of ICSLP, Denver, Colorado