DIRA: Dialectal Arabic Information Retrieval Assistant

Arfath Pasha, Mohammad Al-Badrashiny, Mohamed Altantawy, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth
Center for Computational Learning Systems
Columbia University, New York, NY
Mona Diab
Department of Computer Science
The George Washington University
dira@ccls.columbia.edu

Abstract

DIRA is a query expansion tool that generates search terms in Standard Arabic and/or its dialects when provided with queries in English or Standard Arabic. The retrieval of dialectal Arabic text has recently become necessary due to the increase of dialectal content on social media. DIRA addresses the challenges of retrieving information in Arabic dialects, which have significant linguistic differences from Standard Arabic. To our knowledge, DIRA is the only tool in existence that automatically generates dialect search terms with relevant morphological variations from English or Standard Arabic query terms.

1 Introduction

The Arabic language poses two problems for information retrieval (IR). First, Arabic is morphologically rich, which increases the likelihood of mismatch between words used in queries and words in documents. Much work has been done on addressing this issue in the context of Modern Standard Arabic (MSA), primarily using different methods of stemming and query reformulation (Al-Kharashi and Evens, 1999; Darwish et al., 2005; Habash et al., 2006; Larkey et al., 2007).[1]

Secondly, the Arabic-speaking world displays diglossia, meaning that a standard language, MSA, co-exists with dialects, such as Egyptian Arabic (EGY). The dialects differ from MSA in many dimensions, which limits the effectiveness of using MSA tools to handle the dialects. Relevant to IR are lexical and morphological differences. Lexically, different words may be used to convey the same meaning in different dialects and MSA. Table 1 presents the same set of four words in English, MSA, Egyptian Arabic and Levantine Arabic.[2]

English	MSA	Egyptian	Levantine
to see	رأى *rÂý*	شاف *šAf*	شاف *šAf*
only	فقط *faqaT*	بس *bas*	بس *bas*
table	طاولة *TAwilaħ*	طربيزة *tarabayzaħ*	طاولة *TAwliħ*
wife [of]	زوجة *zawjaħ*	مرات *mirAt*	مرت *mart*
these	هؤلاء *hawˆlA'*	دول *dawl*	هدول *hadawl*

Table 1: Four examples showing lexical variation among Arabic dialects and MSA.

Morphologically, the dialects may use different forms from MSA, e.g., the short phrase 'he writes' appears as يكتب *yaktubu* in MSA , but as بيكتب *biyiktib* in EGY, ديكتب *dayiktib* in Iraqi Arabic and كيكتب *kayiktib* in Moroccan Arabic. The differences between MSA and dialect morphology can be rather large: Habash et al. (2012a) report that over one-third of EGY words cannot be analyzed using an MSA morphological analyzer; and Habash and Rambow (2006) report similar figures for Levantine verbs.

Furthermore, while MSA has a standard orthography, the dialects are not orthographically standardized, which leads to the coexistence of multiple spellings for the same word, e.g., the future marker in EGY may be written as ه *h* or ح *H*. We address this problem in the context of natural language processing of Arabic dialect by proposing a conventional orthography for representing dialec-

[1] For more information on Arabic natural language processing issues, see (Habash, 2010).

[2] Arabic transliteration throughout the paper is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTĎςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ى ħ ة, ý ى.

The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, pages 13–16,
Nagoya, Japan, 14-18 October 2013.

tal Arabic elsewhere (Habash et al., 2012b).

Traditionally, almost all written Arabic was in MSA and not in the dialects. The retrieval of dialectal Arabic text has recently become necessary due to the increase of dialectal content on social media that is not "curated" (i.e., not chosen or edited by professionals). Our tool, DIRA (Dialectal [Arabic] Information Retrieval Assistant), is a query expansion tool that generates search terms, comprising both lexical and morphological variants, in MSA and EGY when provided with queries in English or MSA. No stemming decisions are made as part of DIRA in order to allow its output to be usable by a variety of IR systems with different stemming decisions. While the problem of morphological richness in IR has been addressed before, our DIRA system is, to our knowledge, the only system that addresses the problem of dialectal variation.

In the next three sections, we discuss DIRA's functionality, some of DIRA's implementation details, and two use scenarios.

2 DIRA's Functionality

DIRA is designed to be used as a component in a cross-lingual information retrieval system (Gey and Oard, 2001). Its purpose is to allow English and Arabic speakers to search for MSA and dialectal content using English or MSA queries. For instance, teachers and language learners may use English queries in DIRA to search the web for sentences containing certain MSA or EGY inflected word forms. An Arabic speaker may use MSA queries in DIRA to search for online EGY content.

The interface accepts English or MSA lemmas (citation forms) as input. MSA lemmas can be undiacritized or (partially) diacritized. Depending on user choice, DIRA outputs a set of MSA or EGY inflected forms for each lemma. The expansions are scored and ranked based on their frequency of use in large MSA and EGY corpora. Advanced settings give the users of DIRA the ability to specify weights for different inflectional feature values such as singular number, imperfective aspect, masculine gender, etc. This allows the system to prefer certain feature values that may be used more often in certain types of content. For instance, 1st and 2nd person may be used more often than 3rd person in blog articles while the converse may be true for news related articles. The weight of specific feature-value pairs can also be set to zero, thus eliminating their corresponding inflected forms from the expanded query.

We demonstrate DIRA's utility in a web application that uses Google search as the IR system. In this application, DIRA first translates (if needed) user queries (in English or MSA) and then morphologically expands the lemmas in the target language or dialect. Google's boolean search operators are used to concatenate a user-selected subset of the generated search terms to build the final search query. This final search query is used to perform a Google search for related online material. The demo web application shows the generated search terms as well as the Google search results. See Figures 1 and 2.

The online demo is available at **http://nlp.ldeo.columbia.edu/dira**.

3 DIRA's Implementation

DIRA expansion consists of three stages: lemma translation, morphological generation, and output ranking. First, DIRA translates each input lemma into a set of target lemmas using a trilingual English-MSA-EGY dictionary containing about 70,000 entries (Diab et al., in preparation). Second, DIRA morphologically expands the target lemmas into sets of inflected word forms using a target-language morphological generator (Habash, 2007). For MSA, the generator uses the databases of the BAMA/SAMA morphological analyzer (Buckwalter, 2004; Graff et al., 2009). For EGY, it uses the databases of the CALIMA-ARZ analyzer (Habash et al., 2012a). Since the CALIMA-ARZ analyzer maps a set of common spelling variations to the conventional orthography we use for EGY (Habash et al., 2012b), in generation mode, different spelling variants are produced. This is a desirable feature as it allows us to match more terms. In order to speed up the expansion process, DIRA utilizes a lookup cache created from large MSA and EGY corpora and extended online with new generated forms. Third, DIRA ranks the expansions using a weighted combination of (a) lemma-feature probabilities estimated from large MSA and EGY annotated corpora, and (b) user-provided weights for various feature-value pairs.

The DIRA framework has been designed to be easily extended to other dialects. At a minimum, a dialectal-MSA-English dictionary and a databases for morphological generation are required. Additional optional resources include corpora for the new dialects that can be used to estimate different probabilities.

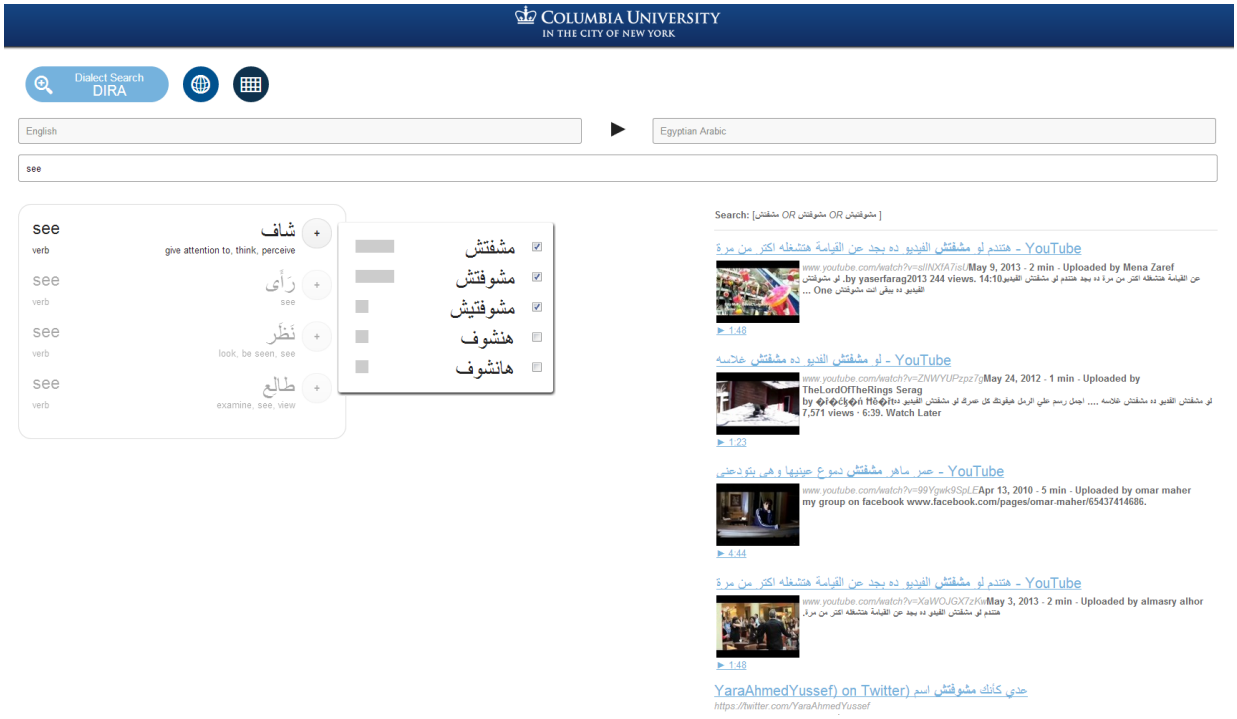

Figure 1: Screenshot of the DIRA demo web application. In this example, the user entered the English query 'see' and requested that the translation and expansion target Egyptian Arabic.

4 Two Use Scenarios

We discuss next two use scenarios. In the first scenario, a teacher of Arabic as a foreign language wishes to use real materials to teach the negation forms in Egyptian Arabic. This scenario is illustrated in Figure 1. She selects English as the input language and Egyptian Arabic as the output language. She chooses to search for the verb "see". The system provides an English gloss for each lemma to help semantically distinguish different lemmas. The teacher can change the lemma choice, but she doesn't because the first lemma is strongly dialectal. To see the available inflected forms for this lemma, she clicks on the plus sign next to the lemma. The online system proposes a maximum of five inflected forms per lemma. The first two are automatically selected by the system and both happen to express morphological negation. She additionally selects the third term, which is also negated.

As soon as the English search query is entered, the system immediatelty returns four lemmas, uses two inflected forms of the top-ranked lemma to construct the search query, and diisplays the results of the search with that seach query. The query and the results of the Google search are shown on the right hand side of the interface. As the user modifies the choice of lemma or inflected forms on the left-hand side of the interface, the query and search results are immediately updated to the right. In Figure 1, we see the search query is the disjunction of the three inflected forms our user selected.

In the second scenario, a native speaker of Arabic who may not know Egyptian Arabic wishes to conduct a search in Egyptian Arabic. This scenario is illustrated in Figure 2. He selects Standard Arabic as the input language and Egyptian Arabic as the output language. He enters the MSA question 'أين مرسي' *Âyn mrsy* 'where is Morsi' as his base query. DIRA expands identifies two possible lemma matches for each term. For the first word, it generates the verbal lemma أيّن *Âyn* 'ionize' and the interrogative particle lemma فين *fyn* 'where'. For the second word, it generates the noun lemma مرسى *mrsý* 'harbor' and the proper noun lemma مرسي *mrsy* 'Morsi'. For both terms, the first lemma is automatically selected. The user deselects the system's automatic choices and clicks on the second reading for each term as these choices fit his intended query. As in the first scenario, after each choice is made, the query terms are adjusted and the search results presented immediately.

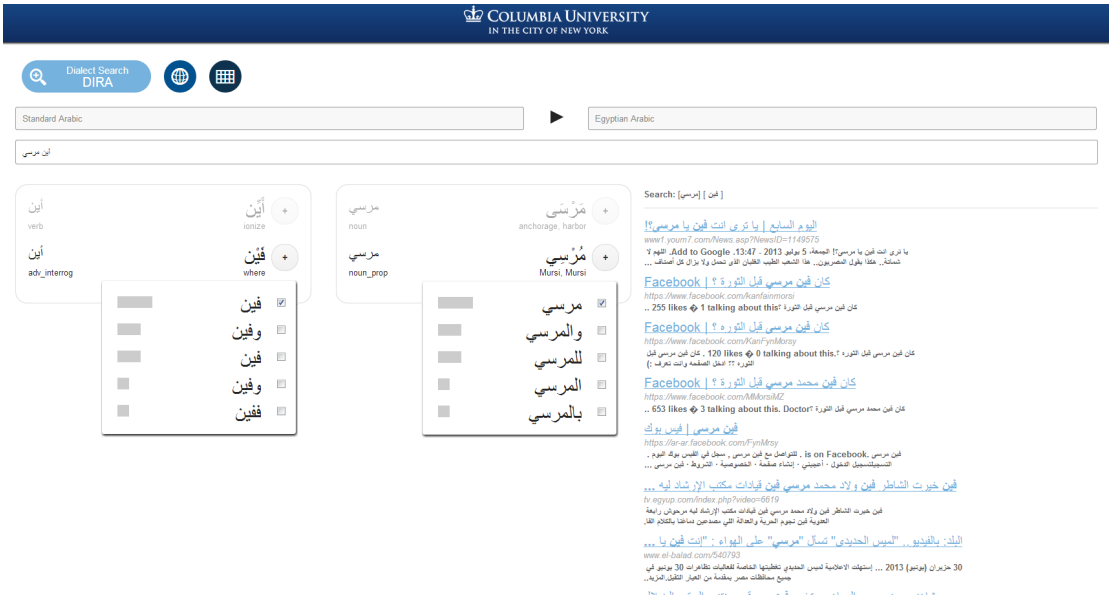

Figure 2: Screenshot of the DIRA demo web application. In this example, the user entered the MSA query ‘أين مرسي’ *Âyn mrsy* ‘where is Morsi’ and requested that the translation and expansion target Egyptian Arabic.

References

Ibrahim A Al-Kharashi and Martha W Evens. 1999. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. *Journal of the American Society for Information Science*, 45(8):548–560.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Kareem Darwish, Hany Hassan, and Ossama Emam. 2005. Examining the effect of improved context sensitive morphology on Arabic information retrieval. *Computational Approaches to Semitic Languages*, 100:25.

Mona Diab, Abdelati Hawwari, Heba Elfardy, Pradeep Dasigi, Mohammad Al-Badrashiny, Ramy Eskander, and Nizar Habash. in preparation. Tharwa: A multi-dialectal multi-lingual machine readable dictionary.

F. Gey and D. Oard. 2001. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. In *The 10th Text Retrieval Conference (TREC-10)*.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of ACL'06*, Sydney, Australia.

Nizar Habash, Clinton Mah, Sabiha Imran, Randy Calistri-Yeh, and Páraic Sheridan. 2006. Design, Construction and Validation of an Arabic-English Conceptual Interlingua for Cross-lingual Information Retrieval. In *LREC-2006*, Genoa, Italy.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

N. Habash, R. Eskander, and A. Hawwari. 2012a. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Nizar Habash, Mona Diab, and Owen Rabmow. 2012b. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Nizar Habash. 2007. Arabic Morphological Representations for Machine Translation. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell, 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, chapter Light Stemming for Arabic Information Retrieval. Springer Netherlands, Kluwer/Springer edition.