DALILA: The Dialectal Arabic Linguistic Learning Assistant

Salam Khalifa, Houda Bouamor[†], Nizar Habash
Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE
[†]Carnegie Mellon University in Qatar, Qatar
{salamkhalifa,nizar.habash}@nyu.edu, hbouamor@cmu.edu

Abstract

Dialectal Arabic (DA) poses serious challenges for Natural Language Processing (NLP). The number and sophistication of tools and datasets in DA are very limited in comparison to Modern Standard Arabic (MSA) and other languages. MSA tools do not effectively model DA which makes the direct use of MSA NLP tools for handling dialects impractical. This is particularly a challenge for the creation of tools to support learning Arabic as a living language on the web, where authentic material can be found in both MSA and DA. In this paper, we present the Dialectal Arabic Linguistic Learning Assistant (DALILA), a Chrome extension that utilizes cutting-edge Arabic dialect NLP research to assist learners and non-native speakers in understanding text written in either MSA or DA. DALILA provides dialectal word analysis and English gloss corresponding to each word.

Keywords: Dialectal Arabic, Computer Assisted Language Learning, Morphology

1. Introduction

Computer-assisted language learning is an area where researchers study how to make use of computers to support learners of a certain language (Chapelle, 2008; Levy and Stockwell, 2013). Many software tools were designed to help learners by providing dictionaries, expert feedback, carefully designed "immersive" experiences, etc. However, language learners see little support online especially with real raw materials in the language they are learning. There is no substitute for a language learner to go online and read and understand text written by native speakers.

In the case of Arabic, the situation is more challenging than English and other European languages. The problem that most learners of Arabic face is that the written language is radically different from the various forms of Dialectal Arabic (DA) spoken throughout the Arab World. Hence, even the most advanced learners of Modern Standard Arabic (MSA) or Moroccan Arabic (to pick a dialect randomly) will find themselves confused on the streets of Cairo as well as on microblogging websites. It is also important to note that even "native" speakers have trouble understanding other dialect vocabulary (Holes, 2004). Different DAs differ phonologically, lexically, morphologically, and syntactically from one another and from MSA (Holes, 1986; Watson, 2007; Brustad, 2000; Habash and Rambow, 2006; Habash et al., 2012; Bouamor et al., 2014). For example, state-of-the-art MSA morphological analyzers have been shown to have only 60% coverage of Levantine Arabic verbs (Habash and Rambow, 2006) and 64% coverage of Egyptian Arabic words (Habash et al., 2012).

While MSA is commonly used in formal written contexts, people are increasingly using DA on social media platforms such as Facebook, Twitter or Youtube to share their stories, opinions, post a comment or give a feedback or comment on a video and interact with the community (Salama et al., 2014; Sadat et al., 2014). In a recent analysis, Huang (2015) showed that 42% of a sample of Facebook Arabic posts are written in DA (most of them in Egyptian).

Using online machine translation systems such as Google Translate to translate the DA sentences into English often produces incorrect or nonsense translations because such systems are trained on MSA-English parallel data. For instance, in the example given in Figure 1., the word مابتفكرش *mAbtfkrš* 'She doesn't think',[1] was considered as an out of vocabulary and was erroneously transliterated instead (*mAptvkrh*). In order to overcome such issues and provide a continuous online/offline assistance, we present the Dialectal Arabic Linguistic Learning Assistant (DALILA),[2] a Chrome extension that utilizes cutting edge NLP tools targeted towards DA to assist learners and non-native speakers in understanding different texts written in either DA (and also MSA). DALILA provides word analysis and English gloss corresponding to each word.

2. Arabic and its Dialects

Arabic processing is challenging for a number of reasons (Habash, 2010). Arabic is both morphologically rich (with many different affixes and clitics) and very ambiguous because it is written with an orthography that omits short vowels. Furthermore, Arabic has a number of variants (MSA and different DAs). While MSA is the shared official language of culture, media and education from Morocco to the Arabian Gulf countries, it is not the native language of any speakers of Arabic. DA is nowadays emerging as the language of informal communication online; in emails, blogs, discussion forums, chats, etc., as the media which is closer to the spoken form of language. A common breakdown of Arabic varieties into dialect groups identifies five major Arabic dialect regions: Egyptian, Gulf, Maghrebi, Levantine and Iraqi (Habash, 2010). There are other minor di-

[1]Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order)
أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
Â b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional symbols: ' ء, Â أ, Ă إ, Ā آ, ŵ ؤ, ŷ ئ, ى ىء, ħ ة, ý ى.

[2]DALILA دليلة *dlylħ* is an Arabic word meaning '*guide[fem]*'.

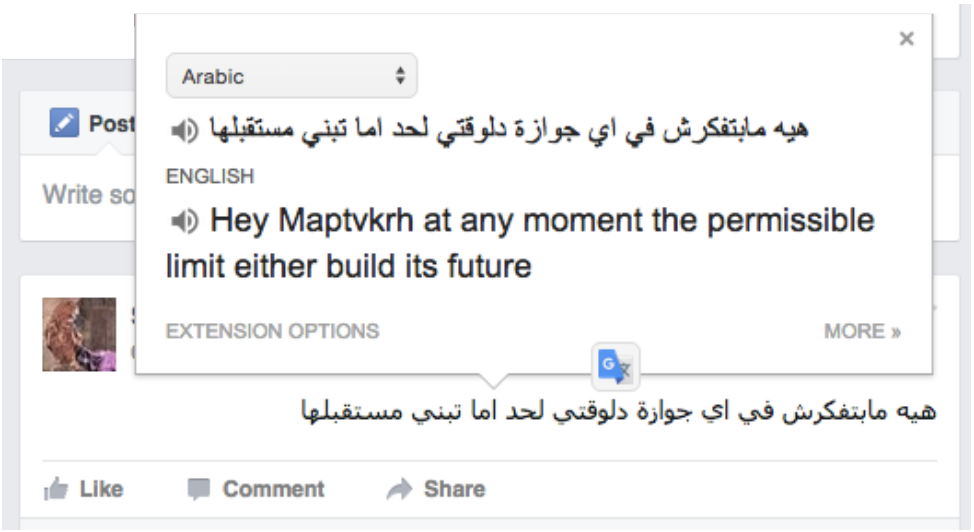

Figure 1: Example of a Facebook post written in Egyptian Arabic and its translation obtained using Google Translate Chrome Extension on October, 21 2015. The correct English translation is 'She won't be thinking about marriage until she establishes herself. [lit. *she not-she-thinks-not in any marriage now until-when she-builds her-future*].'

alects as well and some can be considered a different class on its own, e.g. Yemeni.
Recently, automatic Arabic dialect processing has attracted a considerable amount of research in NLP (Shoufan and Alameri, 2015). Most of these focus on the following:

- Morphological processing tools (Pasha et al., 2014; Habash et al., 2013; Habash et al., 2012)
- Annotation tools (Al-Shargi and Rambow, 2015)
- Resource creation (Zaidan and Callison-Burch, 2011; Bouamor et al., 2014; Diab et al., 2014; Khalifa et al., 2016; Al-Shargi et al., 2016), and
- Developing DA to English machine translation systems (Salloum and Habash, 2011; Zbib et al., 2012; Sajjad et al., 2013; Elfardy et al., 2014; Al-Badrashiny et al., 2014; Huang, 2015)

Our system makes use of state-of-the-art suite of tools developed to perform a morphological, syntactic and semantic analysis of DA and MSA texts.

3. The DALILA tool

In this section we talk about the motivation behind building DALILA, we then present our approach to use current research and available technologies to build a simple tool that is accessible and useful to the general public.

3.1. Motivation

The different challenges of Arabic dialects as discussed above can be hard for both learners and non-native speakers within a diverse community such as social networks. Given that most available resources are developed for MSA such as Google translate, or the embedded Bing translator inside Facebook and Twitter, it is convenient to have a resource that can fulfill the needs of both learners and non-native speakers to understand DA. Given the differences among different DAs, this is also helpful for Arabic native speakers from different dialects.

3.2. Solution

There has been much research done in the area of DA. Such research can be utilized as a backend for tools targeted for the general public. In our work, we opted to develop a browser extension. This will enable the tool to be cross platform, available to all technology users and can be used online and offline. There exist many similar extensions for other languages such as Japanese[3] and Korean.[4] These tools are simple dictionary entries and lookup tables on the backend, and the front end is a simple graphical user interface that shows the answer to the user.

3.3. Ingredients of the Extension

We decided to develop a Google Chrome extension for its simplicity and existing templates and mainly because of the popularity of the Chrome browser.[5] The extension is developed mainly using JavaScript and JSON (JavaScript Object Notation) dictionaries for the actual dictionary entries.

3.3.1. Front End

The front end is the interactive part of the extension. Here, we use the double click on a single word to select it, and then produce a small pop-up displaying the answers. Figure 3. shows an example of the actual interaction taking place.

3.3.2. Backend

The core of the backend is the basic morphological analysis algorithm used in ALMOR (Habash, 2007) (which is based on the algorithm of BAMA (Buckwalter, 2004)). By using the ALMOR algorithm, we have access to any of

[3] https://code.google.com/p/rikaikun/
[4] http://www.toktogi.com/
[5] http://www.w3schools.com/browsers/

Figure 2: (A) Word selected by double clicking. (B) Pop-up balloon with answers.

In this section we

the database in the ALMOR suite of Arabic Morphological databases. In the example in Figure 3., we specifically use the Egyptian Arabic morphological analyzer (CALIMA-ARZ)'s database (Habash et al., 2012). The database is converted to separate JSON dictionary files for suffix, stem and prefix. We also use two simple unigram language models for part-of-speech (POS) and lemmas (3016 and 12823 unigram entries respectively) based on the Penn Arabic Treebank (PATB parts 1,2 and 3) (Maamouri et al., 2004; Maamouri et al., 2006; Maamouri et al., 2009) with the training data split along the recommendations from Diab et al. (2013), which is used in the training of Arabic morphological tagger MADAMIRA (Pasha et al., 2014) that we later compare to.

The morphological analysis algorithm is implemented in JavaScript. The main tasks of the algorithm are:

- Word is segmented into every possible 'prefix+stem+suffix' sequences, including empty prefixes and suffixes.
- Each prefix, stem and suffix is checked against the dictionary to find all possible valid forms.
- The combination prefix+stem+suffix analysis with the highest probability (in terms of POS and lemma choices) is selected.
- The selected analysis is sent to the front end.

3.4. Evaluation

We perform a manual quantitative analysis on 1,000 words of Egyptian Arabic taken from the online blog عايزة اتجوز *ʕAyzħ Atjwz* 'I wanna get married'.[6] For every word, we evaluate on the correctness of both the top analysis from MADAMIRA-EGY and the analysis produced by DALILA in terms of POS and English gloss (as a substitute for lemma choice). The results are in Table 1. The results of MADAMIRA-EGY are comparable to the system evaluation by Pasha et al. (2014) where POS accuracy was 92.4%. However, the lemma accuracy reported by Pasha et al. (2014) is higher than our English gloss accuracy (87.8%). This is likely due to the difference in the genre of the training data. The accuracy results of DALILA are lower than MADAMIRA-EGY by 4.8% and 6.8% absolute in terms of POS and English gloss, respectively. This is not unexpected given the simpler disambiguation algorithm used in DALILA. The difference is small given the tradeoff of system complexity.

Feature	DALILA	MADAMIRA-EGY
POS	88.50	93.30
English Gloss	72.70	79.50

Table 1: Results on quantitative analysis of the output of DALILA and MADAMIRA-EGY

4. Conclusion and Future Work

We presented DALILA, a Chrome extension that utilizes cutting-edge Arabic dialect NLP research to assist learners and non-native speakers in understanding text written in either MSA or DA. DALILA provides dialectal word analysis and English gloss corresponding to each word.

In the future, we plan to cover more dialects as their databases become available. We also plan to use a more advanced algorithm that would take context into account as the current algorithm only works on single words. Finally, we plan to extend the tool's front end to allow different styles of presenting information back to the user including highlighting ambiguity and pronunciation details.

[6] http://wanna-b-a-bride.blogspot.ae

5. Acknowledgments

We would like to thank all the members of the team that developed the initial version of DALILA during the New York University Abu Dhabi Hackathon for Social Good in the Arab World (2015). The team members are Ayman Awartani, Clair Seager, Ayah Soufan, Maeda Hanafi, Hiba Bejaoui, Fatima Al Neyadi, Sara Al Kendi and the authors of this paper.

6. Bibliographical References

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic Transliteration of Romanized Dialectal Arabic. *CoNLL-2014*, page 30.

Al-Shargi, F. and Rambow, O. (2015). DIWAN: A Dialectal Word Annotation Tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, Beijing, China.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). A Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and San'ani Yemeni Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland.

Brustad, K. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

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

Chapelle, C. A. (2008). *Computer Assisted Language Learning*. Wiley Online Library.

Diab, M., Habash, N., Rambow, O., and Roth, R. (2013). LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). Aida: Identifying code switching in informal arabic text. *EMNLP 2014*, pages 94–101.

Habash, N. and Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of ACL*, pages 681–688, Sydney, Australia.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Eskander, R., and Hawwari, A. (2012). A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada, June. Association for Computational Linguistics.

Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of NAACL-HLT*, pages 426–432, Atlanta, Georgia, June.

Habash, N. (2007). Arabic Morphological Representations for Machine Translation. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Holes, C. (1986). Variation in the Morphophonology of Arabic Dialects. *Transactions of the Philological Society*, 84(1):167–190.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press. Revised Edition.

Huang, F. (2015). Improved Arabic Dialect Classification with Social Media Data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2118–2126, Lisbon, Portugal.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Levy, M. and Stockwell, G. (2013). *CALL Dimensions: Options and Issues in Computer-assisted Language Learning*. Routledge.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Maamouri, M., Bies, A., and Kulick, S. (2006). Diacritization: A challenge to arabic treebank annotation and parsing. In *Proceedings of the Conference of the Machine Translation SIG of the British Computer Society*.

Maamouri, M., Bies, A., and Kulick, S. (2009). Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating dialectal arabic to english. In *Proceedings of ACL*, Sofia, Bulgaria.

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014). YouDACC: the Youtube Dialectal Arabic Com-

ment Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland.

Salloum, W. and Habash, N. (2011). Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Shoufan, A. and Alameri, S. (2015). Natural Language Processing for Dialectical Arabic: A Survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China.

Watson, J. C. (2007). *The Phonology and Morphology of Arabic*. Oxford University Press.

Zaidan, O. F. and Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the Association for Computational Linguistics*, Portland, Oregon, USA.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine Translation of Arabic Dialects. In *Proceedings of NAACL-HLT*, Montréal, Canada.