Elissa: A Dialectal to Standard Arabic Machine Translation System

Wael Salloum and Nizar Habash
Center for Computational Learning Systems
Columbia University
{wael,habash}@ccls.columbia.edu

Abstract
Modern Standard Arabic (MSA) has a wealth of natural language processing (NLP) tools and resources. In comparison, resources for dialectal Arabic (DA), the unstandardized spoken varieties of Arabic, are still lacking. We present Elissa , a machine translation (MT) system from DA to MSA. Elissa (version 1.0) employs a rule-based approach that relies on morphological analysis, morphological transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences. Elissa can be employed as a general preprocessor for dialectal Arabic when using MSA NLP tools.

إليسَّا: نظَامٌ حَاسوبيٌّ للتَّرجمةِ الآليَّةِ من العَامِّيَّاتِ العربيَّةِ إِلَى العربيَّةِ الفُصحَى

توجدُ أدَوَاتٌ ومَوَاردُ كثيرةٌ لمعَالجةِ اللّغةِ العربيَّةِ الفُصحَى حَاسوبيًّا بينَمَا لَا تتوفَّرُ أدوَاتٌ ومَوَاردُ مُمَاثلةٌ لمُعَالجة العَاميَّاتِ العربيَّة، وهي النُّسخُ المحكيَّةُ غيرُ القيَاسيَّةِ من اللُّغة العربيَّة. سنقدمُ في بحثنَا هذَا إليسَّا، وهي نظَامٌ حَاسوبيٌّ يقومُ بالتَّرجمةِ الآليَّةِ من العَاميَّاتِ العربيَّة إِلَى العربيَّة الفُصحَى. تعتمدُ إليسَّا حلًّا مبنيًّا علَى القَوَاعدِ، تستخدِمُ فيه التَّحليلَ الصَّرفيَّ للكلمةِ ومجموعةً من قوَاعدِ التَّرجمةِ ومعَاجمَ عَامِّيَّةً لإنشَاءِ مرَادفَاتٍ وترجمَاتٍ للكلمَاتِ العَامِّيَّةِ، إِضَافةً إِلَى نَمَاذِجَ لُغويَّةٍ لاختيَارِ الجملةِ الفُصحَى الأَفضل طلَاقةً بينَ جميعِ الجُمَلِ المُمكنة. يمكنُ استخدَامُ إليسَّا لمُعَالجة العَاميَّاتِ العربيَّةِ قبلَ استخدَامِ أدَوَاتٍ مُعَدَّةٍ للّغةِ العربيَّةِ الفُصحَى عليهَا.

Keywords: Dialectal Arabic, Arabic Natural Language Processing, Machine Translation, Rule-Based Machine Translation, Morphology.

Keywords in L_2:

العَاميَّات العربيَّة، مُعَالجة اللغة العربية حَاسوبيًّا، الترجمة الآلية، الترجمة الآلية المعتمدة عَلَى القَوَاعد، عِلمُ الصَّرف.

Proceedings of COLING 2012: Demonstration Papers, pages 385–392,
COLING 2012, Mumbai, December 2012.

1 Introduction

Much work has been done on Modern Standard Arabic (MSA) natural language processing (NLP) and machine translation (MT). MSA has a wealth of resources in terms of morphological analyzers, disambiguation systems, annotated data, and parallel corpora. In comparison, research on dialectal Arabic (DA), the unstandardized spoken varieties of Arabic, is still lacking in NLP in general and MT in particular. In this paper we present Elissa, our DA-to-MSA MT system.[1] To process Arabic dialects with available MSA tools, researchers have used different techniques to map DA words to their MSA equivalents (Chiang et al., 2006; Sawaf, 2010; Salloum and Habash, 2011). Having a publicly available tool that translates DAs to MSA can help researchers extend their MSA resources and tools to cover different Arabic dialects. Elissa currently handles Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic.

The Elissa approach can be summarized as follows. First, Elissa uses different techniques to identify dialectal words in a source sentence. Then, Elissa produces MSA paraphrases for the selected words using a rule-based component that depends on the existence of a dialectal morphological analyzer, a list of morphological transfer rules, and DA-MSA dictionaries. The resulting MSA is in a lattice form that we pass to a language model for n-best decoding which selects the best MSA translations.

2 Challenges for Processing Arabic and its Dialects

Contemporary Arabic is in fact a collection of varieties: MSA, the official language of the Arab World, which has a standard orthography and is used in formal settings; and DAs, the commonly used informal native varieties, which have no standard orthographies but have an increasing presence on the web. Arabic, in general, is a morphologically complex language which has rich inflectional morphology, expressed both templatically and affixationally, and several classes of attachable clitics. For example, the Arabic word وسيكتبونها *w+s+y-ktb-wn+hA*[2] 'and they will write it' has two proclitics (+و *w+* 'and' and +س *s+* 'will'), one prefix -ي *y-* '3rd person', one suffix -ون *-wn* 'masculine plural' and one pronominal enclitic +ها *+hA* 'it/her'. DAs differ from MSA phonologically, morphologically and to some lesser degree syntactically. The morphological differences are most noticeably expressed in the use of clitics and affixes that do not exist in MSA. For instance, the Levantine Arabic equivalent of the MSA example above is وحيكتبوها *w+H+y-ktb-w+hA* 'and they will write it'. The optionality of vocalic diacritics helps hide some of the differences resulting from vowel changes; compare the diacritized forms: Levantine *wHayikitbuwhA* and MSA *wasayaktubuwnahA*.

All of the NLP challenges of MSA (e.g., optional diacritics and spelling inconsistency) are shared by DA. However, the lack of standard orthographies for the dialects and their numerous varieties pose new challenges. Additionally, DAs are rather impoverished in terms of available tools and resources compared to MSA, e.g., there is very little parallel DA-English corpora and almost no MSA-DA parallel corpora. The number and sophistication of morphological analysis and disambiguation tools in DA is very limited in comparison to MSA (Duh and Kirchhoff, 2005; Habash and Rambow, 2006; Abo Bakr et al., 2008; Salloum and Habash, 2011; Habash et al., 2012). MSA tools cannot be effectively used to handle DA, e.g., Habash and Rambow (2006) report that over one-third of Levantine verbs cannot be analyzed using an MSA morphological analyzer.

[1] Elissa's home page can be found at http://nlp.ldeo.columbia.edu/elissa/.

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

3 Related Work

Much work has been done in the context of MSA NLP (Habash, 2010). In contrast, research on DA NLP is still in its early stages: (Kilany et al., 2002; Kirchhoff et al., 2003; Duh and Kirchhoff, 2005; Habash and Rambow, 2006; Chiang et al., 2006; Habash et al., 2012; Elfardy and Diab, 2012). Several researchers have explored the idea of exploiting existing MSA rich resources to build tools for DA NLP (Chiang et al., 2006). Such approaches typically expect the presence of tools/resources to relate DA words to their MSA variants or translations. Given that DA and MSA do not have much in terms of parallel corpora, rule-based methods to translate DA-to-MSA or other methods to collect word-pair lists have been explored (Abo Bakr et al., 2008; Sawaf, 2010; Salloum and Habash, 2011). Using closely related languages has been shown to improve MT quality when resources are limited (Hajič et al., 2000; Zhang, 1998). This use of "resource-rich" related languages is a specific variant of the more general approach of using pivot/bridge languages (Utiyama and Isahara, 2007; Kumar et al., 2007). Sawaf (2010) built a hybrid DA-English MT system that uses an MSA pivoting approach. In his approach, DA is normalized into MSA using character-based DA normalization rules, a DA morphological analyzer, a DA normalization decoder that relies on language models, and a lexicon. Similarly, we use some character normalization rules, a DA morphological analyzer, and DA-MSA dictionaries. In contrast, we use hand-written morphological transfer rules that focuses on translating DA morphemes and lemmas to their MSA equivalents. We also provide our system to be used by other researchers. In previous work, we built a rule-based DA-MSA system to improve DA-to-English MT (Salloum and Habash, 2011). We applied our approach to ATB-tokenized Arabic. Our DA-MSA transfer component used feature transfer rules only. We did not use a language model to pick the best path; instead we kept the ambiguity in the lattice and passed it to our SMT system. In this work, we run Elissa on untokenized Arabic, we use feature, lemma, and surface form transfer rules, and we pick the best path of the generated MSA lattice through a language model. In this paper, we do not evaluate Elissa. We reserve the evaluation to a future publication.

4 Elissa

Elissa is a Dialectal Arabic to Modern Standard Arabic Translation System. It is available for use by other researchers. In Elissa 1.0 (the version we present in this paper), we use a rule-based approach (with some statistical components) that relies on the existence of a dialectal morphological analyzer, a list of hand-written transfer rules, and DA-MSA dictionaries to create a mapping of DA to MSA words and construct a lattice of possible sentences. Elissa uses a language model to rank and select the generated sentences.

4.1 Elissa Input/Output

Elissa supports input encoding in Unicode (UTF-8) or Buckwalter transliteration (Buckwalter, 2004). Elissa supports untokenized (i.e., raw) input only. The output of Elissa can be encoded also in Unicode or Buckwalter transliteration. Elissa supports the following types of output:

1. **Top-1 sentence.** Elissa uses an untokenized MSA language model to rank the paths in the MSA translation output lattice. In this output format, Elissa selects the top-1 choice (the best path) from the ranked lattice.

2. **N-Best sentences.** Using the untokenized MSA language model, Elissa selects the top 'n' sentences from the ranked lattice. The integer 'n' is configurable.

3. **Map file.** Elissa outputs a file that contains a list of entries of the format: source-word, weight, target-phrase. The weight is calculated in the transfer component not by the language model.

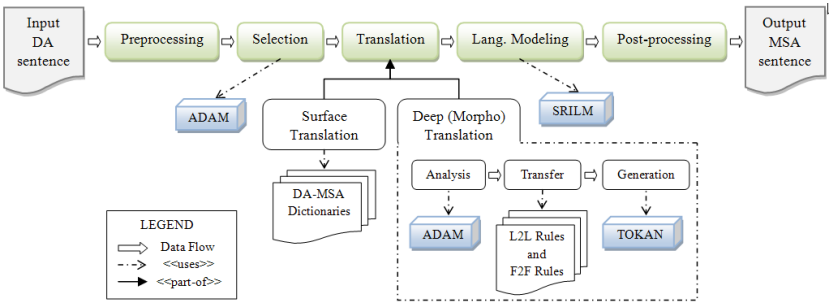

Figure 1: This diagram highlights the different steps inside Elissa and some of its third-party dependencies. ADAM is an Analyzer for Dialectal Arabic Morphology (Salloum and Habash, 2011). TOKAN is a general tokenizer for Arabic (Habash, 2007). SRILM is SRI Language Modeling Toolkit (Stolcke, 2002). ADAM and TOKAN are packaged with Elissa.

This variety of output types makes it easy to plug Elissa with other systems and to use it as a dialectal Arabic preprocessing tool for other MSA systems, e.g., MADA (Habash and Rambow, 2005) or AMIRA (Diab et al., 2007).

4.2 Approach

Our approach, illustrated in Figure 1, consists of three major steps proceeded by a *preprocessing* step, that prepares the input text to be handled (e.g., UTF-8 cleaning), and succeeded by a *post-processing* step, that produces the output in the desired form (e.g., top-1 choice in Buckwalter transliteration). The three major steps are:

1. **Selection.** Identify the words to handle, e.g., dialectal or OOV words.

2. **Translation.** Provide MSA paraphrases of the selected words to form an MSA lattice.

3. **Language Modeling.** Pick the n-best fluent sentences from the generated MSA lattice according to a language model.

4.2.1 Selection

In the first step, Elissa decides which words to paraphrase and which words to leave as is. It provides different alternative settings for selection, and can be configured to use different subsets of them:

1. **User Token-based Selection.** The user can mark specific words for selection using the tag '/DIA' after the word. This allows dialect identification tools, such as AIDA (Elfardy and Diab, 2012), to be integrated with Elissa.

2. **User Type-based Selection.** The user can specify a list of words to select what is listed in it (OOV) or what is not listed in it (INVs – in-vocabulary).

3. **Dialectal Morphology Word Selection.** Elissa uses ADAM (Salloum and Habash, 2011) to select two types of dialectal words: words that have DA analyses only or DA/MSA analyses.

4. **Dictionary-based Selection.** Elissa selects words that exist in our DA-MSA dictionaries.

5. **All.** Elissa selects every word in an input sentence.

4.2.2 Translation

In this step, Elissa translates the selected words to their MSA equivalent paraphrases. These paraphrases are then used to form an MSA lattice. The translation step has two types: *surface translation* and *deep (morphological) translation*. The surface translation depends on DA-to-MSA dictionaries to map a selected DA word directly to its MSA paraphrases. We use the Tharwa dictionary (Diab et al., 2013) and other dictionaries that we created. The morphological translation uses the classic rule-based machine translation flow: analysis, transfer and generation.

1. **Morphological Analysis** produces a set of alternative analyses for each word.

2. **Morphological Transfer** maps each analysis into one or more target analyses.

3. **Morphological Generation** generates the surface forms of the target analyses.

Morphological Analysis. In this step, we use a dialectal morphological analyzer, ADAM, (Salloum and Habash, 2011). ADAM provides Elissa with a set of analyses for each dialectal word in the form of lemma and features. These analyses will be processed in the next step, Transfer.

Morphological Transfer. In the transfer step, we map ADAM's dialectal analyses to MSA analyses. This step is implemented using a set of transfer rules (TRs) that operate on the lemma and feature representation produced by ADAM. These TRs can change clitics, features or lemma, and even split up the dialectal word into multiple MSA word analyses. Crucially the input and output of this step are both in the lemma and feature representation. A particular analysis may trigger more than one rule resulting in multiple paraphrases. This only adds to the fan-out which started with the original dialectal word having multiple analyses.

Elissa uses two types of TRs: lemma-to-lemma (L2L) TRs and features-to-features (F2F) TRs. L2L TRs simply change the dialectal lemma to an MSA lemma. The mapping is provided in the DA-MSA dictionaries we use. On the other hand, F2F TRs are more complicated and were written by experts. These rules work together to handle complex transformations such as mapping the DA circumfix negation to a separate word in MSA and adjusting for verb aspect difference. The following is an example illustrating the various rules working together: Elissa creates the MSA analysis for ولم يذهبوا إليها *wlm yðhbwA ĂlyhA* 'And they did not go to it – lit. and+did+not they+go to+it' starting with the DA analysis for ومارحولا *wmArAHwlA* 'lit. and+not+went+they+to+it'.

Morphological Generation. In this step, we generate Arabic words from all analyses produced by the previous steps. The generation is done using the general tokenizer/generator TOKAN (Habash, 2007) to produce the surface form words. Although TOKAN can accommodate generation in specific tokenizations, in the work we report here we generate only in untokenized form. Any subsequent tokenization is done in a post-processing step (see Section 4.1). The various generated forms are used to construct the map files and word lattices. The lattices are then input to the language modeling step presented next.

4.2.3 Language Modeling

The language model (LM) component uses SRILM (Stolcke, 2002) lattice-tool for weight assignment and n-best decoding. Elissa comes with a default 5-gram LM file (trained on ∼200M untokenized

Arabic words) and default configurations; however, users can change the default configurations and even specify their own LM file.

5 Example

DA source	بهالحالة ماحيكتبولو شي عحيط صفحتو لأنو ماخبرهن يوم اللي وصل عالبلد. *bhAlHAlħ mAHyktbwlw šy EHyT SfHtw lÂnw mAxbrhn ywm Ally wSl EAlbld.*
Human Reference	In this case, they will not write on his page wall because he did not tell them the day he arrived to the country.
Google Translate	Bhalhalh Mahiketbolo Shi Ahat Cefhto to Anu Mabrhen day who arrived Aalbuld.
Human	في هذه الحالة لن يكتبوا له شيئًا على حائط صفحته لأنه لم يخبرهم يوم وصل إلى البلد.
DA-to-MSA	*fy hðh AlHAlħ ln yktbwA lh šyŷA ElY HAŷT SfHth lÂnh lm yxbrhm ywm wSl ĂlY Albld.*
Google Translate	In this case it would not write him something on the wall yet because he did not tell them day arrived in the country.
Elissa	في هذه الحالة لن يكتبوا شيء علي حائط صفحته لانه لم يخبرهم يوم الذي وصل الي البلد.
DA-to-MSA	*fy hðh AlHAlħ ln yktbwA šy' Ely HAŷT SfHth lAnh lm yxbrhm ywm Alðy wSl Aly Albld.*
Google Translate	In this case it would not write something on the wall yet because he did not tell them the day arrived in the country.

Table 1: An illustrative example for DA-to-English MT by pivoting (bridging) on MSA. Elissa's Arabic output is Alif/Ya normalized (Habash, 2010).

Table 1 shows a illustrative example of how pivoting on MSA can dramatically improve the translation quality of a statistical MT system that is trained on mostly MSA-to-English parallel corpora. In this example, we use Google Translate Arabic-English SMT system. The table is divided into three parts. The first part shows a dialectal (Levantine) sentence, its reference translation to English, and its Google Translate translation. The Google Translate translation clearly struggles with most of the dialectal words, which were probably unseen in the training data (i.e., out-of-vocabulary – OOV) and were considered proper nouns (transliterated and capitalized). The lack of DA-English parallel corpora suggests pivoting on MSA can improve the translation quality. In the second part of the table, we show a human MSA translation of the DA sentence above and its Google Translate translation. We see that the results are quite promising. The goal of Elissa is to model this DA-MSA translation automatically. In the third part of the table, we present Elissa's output on the dialectal sentence and its Google Translate translation. The produced MSA is not perfect, but is clearly an improvement over doing nothing as far as usability for MT into English.

Future Work

In the future, we plan to extend Elissa's coverage of phenomena in the handled dialects and to new dialects. We also plan to automatically learn additional rules from limited available data (DA-MSA or DA-English). We are interested in studying how our approach can be combined with solutions that simply add more dialectal training data (Zbib et al., 2012) since the two directions are complementary in how they address linguistic normalization and domain coverage.

Acknowledgment

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

Abo Bakr, H., Shaalan, K., and Ziedan, I. (2008). A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Buckwalter, T. (2004). Buckwalter arabic morphological analyzer version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Diab, M., Hacioglu, K., and Jurafsky, D. (2007). *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, chapter Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking. Springer.

Diab, M., Hawwari, A., Elfardy, H., Dasigi, P., Al-Badrashiny, M., Eskander, R., and Habash, N. (Forthcoming – 2013). Tharwa: A multi-dialectal multi-lingual machine readable dictionary.

Duh, K. and Kirchhoff, K. (2005). POS tagging of dialectal Arabic: a minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Semitic '05, pages 55–62, Ann Arbor, Michigan.

Elfardy, H. and Diab, M. (2012). Token level identification of linguistic code switching. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING), IIT Mumbai, India.*

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

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

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.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Habash, N. and Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia.

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

Hajič, J., Hric, J., and Kubon, V. (2000). Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'2000)*, pages 7–12, Seattle.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D. (2003). Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns Hopkins Summer Workshop. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China.

Kumar, S., Och, F. J., and Macherey, W. (2007). Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic.

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.

Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.

Stolcke, S. (2002). Tokenization, morphological analysis, and part-of-speech tagging for arabic in one fell swoop. In *In Proceedings of ICSLP 2002*, volume 2, pages 901–904.

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R. M., Makhoul, J., Zaidan, O., and Callison-Burch, C. (2012). Machine translation of arabic dialects. In *HLT-NAACL*, pages 49–59.

Zhang, X. (1998). Dialect MT: a case study between Cantonese and Mandarin. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 1460–1464, Montreal, Canada.