Translating Dialectal Arabic to English

Hassan Sajjad, Kareem Darwish
Qatar Computing Research Institute
Qatar Foundation
{hsajjad,kdarwish}@qf.org.qa

Yonatan Belinkov
CSAIL
Massachusetts Institute of Technology
belinkov@mit.edu

Abstract

We present a dialectal Egyptian Arabic to English statistical machine translation system that leverages dialectal to Modern Standard Arabic (MSA) adaptation. In contrast to previous work, we first narrow down the gap between Egyptian and MSA by applying an automatic character-level transformational model that changes Egyptian to EG', which looks similar to MSA. The transformations include morphological, phonological and spelling changes. The transformation reduces the out-of-vocabulary (OOV) words from 5.2% to 2.6% and gives a gain of 1.87 BLEU points. Further, adapting large MSA/English parallel data increases the lexical coverage, reduces OOVs to 0.7% and leads to an absolute BLEU improvement of 2.73 points. We plan to publicly release the Egyptian/MSA word pairs used for training the conversion model.

1 Introduction

Modern Standard Arabic (MSA) is the lingua franca for the Arab world. Arabic speakers generally use dialects in daily interactions. There are 6 dominant dialects, namely Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni[1]. The dialects may differ in vocabulary, morphology, syntax, and spelling from MSA and most lack spelling conventions.

Different dialects often make different lexical choices to express concepts. For example, the concept corresponding to "Oryd" أريد ("I want") is expressed as "EAwz" عاوز in Egyptian, "Abgy" ابغي in Gulf, "Aby" ابي in Iraqi, and "bdy" بدي in Levantine[2]. Often, words have different or opposite meanings in different dialects.

Arabic dialects may differ morphologically from MSA. For example, Egyptian Arabic uses a negation construct similar to the French "ne pas" negation construct. The Egyptian word "mlEbt$" ملعبتش (or alternatively spelled مالعبتش) ("I did not play") is composed of "m+lEbt+$".

The pronunciations of letters often differ from one dialect to another. For example, the letter "q" ق is typically pronounced in MSA as an unvoiced uvular stop (as the "q" in "quote"), but as a glottal stop in Egyptian and Levantine (like "A" in "Alpine") and a voiced velar stop in the Gulf (like "g" in "gavel"). Differing pronunciations often reflect on spelling.

Social media platforms allowed people to express themselves more freely in writing. Although MSA is used in formal writing, dialects are increasingly being used on social media sites. Some notable trends on social platforms include (Darwish et al., 2012):
- Mixed language texts where bilingual (or multilingual) users code switch between Arabic and English (or Arabic and French). In the example "wSlny mrsy" وصلني مرسي ("got it thank you"), "thank you" is the transliterated French word "merci".
– The use of phonetic transcription to match dialectal pronunciation. For example, "Sdq" صدق ("truth") is often written as "Sj" صج in Gulf dialect.
– Creative spellings, spelling mistakes, and word elongations are ubiquitous in social text.

[1] http://en.wikipedia.org/wiki/Varieties_of_Arabic

[2] All transliterations follow the Buckwalter scheme

– The use of new words like "lol" لول ("LOL").
– The attachment of new meanings to words such as using "THn" طحن to mean "very" while it means "grinding" in MSA.

The Egyptian dialect has the largest number of speakers and is the most commonly understood dialect in the Arab world. In this work, we focused on translating dialectal Egyptian to English using Egyptian to MSA adaptation. Unlike previous work, we first narrowed the gap between Egyptian and MSA using character-level transformations and word n-gram models that handle spelling mistakes, phonological variations, and morphological transformations. Later, we applied an adaptation method to incorporate MSA/English parallel data.

The contributions of this paper are as follows:
– We trained an Egyptian/MSA transformation model to make Egyptian look similar to MSA. We will publicly release the training data.
– We built a phrasal Machine Translation (MT) system on adapted Egyptian/English parallel data, which outperformed a non-adapted baseline by 1.87 BLEU points.
– We used phrase-table merging (Nakov and Ng, 2009) to utilize MSA/English parallel data with the available in-domain parallel data.

2 Previous Work

Our work is related to research on MT from a resource poor language (to other languages) by pivoting on a closely related resource rich language. This can be done by either translating between the related languages using word-level translation, character level transformations, and language specific rules (Durrani et al., 2010; Hajič et al., 2000; Nakov and Tiedemann, 2012), or by concatenating the parallel data for both languages (Nakov and Ng, 2009). These translation methods generally require parallel data, for which hardly any exists between dialects and MSA. Instead of translating between a dialect and MSA, we tried to narrow down the lexical, morphological and phonetic gap between them using a character-level conversion model, which we trained on a small set of parallel dialect/MSA word pairs.

In the context of Arabic dialects[3], most previous work focused on converting dialects to MSA and vice versa to improve the processing of dialects (Sawaf, 2010; Chiang et al., 2006; Mohamed et al., 2012; Utiyama and Isahara, 2008). Sawaf (2010) proposed a dialect to MSA normalization that used character-level rules and morphological analysis. Salloum and Habash (2011) also used a rule-based method to generate MSA paraphrases of dialectal out-of-vocabulary (OOV) and low frequency words. Instead of rules, we automatically learnt character mappings from dialect/MSA word pairs.

Zbib et al. (2012) explored several methods for dialect/English MT. Their best Egyptian/English system was trained on dialect/English parallel data. They used two language models built from the English GigaWord corpus and from a large web crawl. Their best system outperformed manually translating Egyptian to MSA then translating using an MSA/English system. In contrast, we showed that training on in-domain dialectal data irrespective of its small size is better than training on large MSA/English data. Our LM experiments also affirmed the importance of in-domain English LMs. We also showed that a conversion does not imply a straight forward usage of MSA resources and there is a need for adaptation which we fulfilled using phrase-table merging (Nakov and Ng, 2009).

2.1 Baseline

We constructed baselines that were based on the following training data:
- An Egyptian/English parallel corpus consisting of ≈38k sentences, which is part of the LDC2012T09 corpus (Zbib et al., 2012). We randomly divided it into 32k sentences for training, 2k for development and 4k for testing. We henceforth refer to this corpus as EG and the English part of it as EG_{en}. We did not have access to the training/test splits of Zbib et al. (2012) to directly compare to their results.
- An MSA/English parallel corpus consisting of 200k sentences from LDC[4]. We refer to this corpus as the AR corpus.

For language modeling, we used either EG_{en} or the English side of the AR corpus plus the English side of NIST12 training data and English GigaWord v5. We refer to this corpus as *GW*.

We tokenized Egyptian and Arabic accord-

[3]Due to space limitations, we restrict discussion to work on dialects only.

[4]Arabic News (LDC2004T17), eTIRR (LDC2004E72), and parallel corpora the GALE program

	Train	LM	BLEU	OOV
B1	AR	GW	7.48	6.7
B2	EG	GW	12.82	5.2
B3	EG	EG_{en}	13.94	5.2
B4	EG	$EG_{en}GW$	**14.23**	5.2

Table 1: Baseline results using the EG and AR training sets with GW and EG_{en} corpora for LM training

ing to the ATB tokenization scheme using the MADA+TOKAN morphological analyzer and tokenizer v3.1 (Roth et al., 2008). Word elongations were already fixed in the corpus. We word-aligned the parallel data using GIZA++ (Och and Ney, 2003), and symmetrized the alignments using grow-diag-final-and heuristic (Koehn et al., 2003). We trained a phrasal MT system (Koehn et al., 2003). We built five-gram LMs using KenLM with modified Kneser-Ney smoothing (Heafield, 2011). In case of more than one LMs, we tuned their weights on a development set using Minimum Error Rate Training (Och and Ney, 2003).

We built several baseline systems as follows:
– *B1* used AR for training a translation model and *GW* for LM.
– *B2-B4* systems used identical training data, namely *EG*, with the *GW*, EG_{en}, or both for *B2*, *B3*, and *B4* respectively for language modeling.

Table 1 reports the baseline results. The system trained on AR (*B1*) performed poorly compared to the one trained on EG (*B2*) with a 6.75 BLEU points difference. This highlights the difference between MSA and Egyptian. Using EG data for training both the translation and language models was effective. *B4* used two LMs and yielded the best results. For later comparison, we only use the *B4* baseline.

3 Proposed Methods

3.1 Egyptian to EG' Conversion

As mentioned previously, dialects differ from MSA in vocabulary, morphology, and phonology. Dialectal spelling often follows dialectal pronunciation, and dialects lack standard spelling conventions. To address the vocabulary problem, we used the EG corpus for training.

To address the spelling and morphological differences, we trained a character-level mapping model to generate MSA words from dialectal ones using character transformations. To train the model, we extracted the most frequent words from a dialectal Egyptian corpus, which had 12,527 news comments (containing 327k words) from Al-Youm Al-Sabe news site (Zaidan and Callison-Burch, 2011) and translated them to their equivalent MSA words. We hired a professional translator, who generated one or more translations of the most frequent 5,581 words into MSA. Out of these word pairs, 4,162 involved character-level transformations due to phonological, morphological, or spelling changes. We aligned the translated pairs at character level using GIZA++ and Moses in the manner described in Section 2.1. As in the baseline of Kahki et al. (2011), given a source word, we produced all of its possible segmentations along with their associated character-level mappings. We restricted individual source character sequences to be 3 characters at most. We retained all mapping sequences leading to valid words in a large lexicon. We built the lexicon from a set of 234,638 Aljazeera articles[5] that span a 10 year period and contain 254M tokens. Spelling mistakes in Aljazeera articles were very infrequent. We sorted the candidates by the product of the constituent mapping probabilities and kept the top 10 candidates. Then we used a trigram LM that we built from the aforementioned Aljazeera articles to pick the most likely candidate in context. We simply multiplied the character-level transformation probability with the LM probability – giving them equal weight. Since Egyptian has a "ne pas" like negation construct that involves putting a "م" and "ش" at the beginning and end of verbs, we handled words that had negation by removing these two letters, then applying our character transformation, and lastly adding the negation article "lA" لا before the verb. We converted the EG train, tune, and test parts. We refer to the converted corpus as EG'.

As an example, our system transformed بس اللّي بيحصلهم ميعجبش حد ("what is happening to them does not please anyone") to بس الذي يحصل لهم لا يعجب حد. Transforming اللّي to الذي involved a spelling correction. The transformation of بيحصلهم to يحصل لهم involved a morphological change and word splitting. The change of ميعجبش to لا يعجب involved morphologically transforming a negation construct.

[5] http://www.aljazeera.net

3.2 Combining AR and EG'

The aforementioned conversion generated a language that is close, but not identical, to MSA. In order to maximize the gain using both parallel corpora, we used the phrase merging technique described in Nakov and Ng (2009) to merge the phrase tables generated from the AR and EG' corpora. If a phrase occurred in both phrase tables, we adopted one of the following three solutions:

- Only added the phrase with its translations and their probabilities from the AR phrase table. This assumed AR alignments to be more reliable.
- Only added the phrase with its translations and their probabilities from the EG' phrase table. This assumed EG' alignments to be more reliable.
- Added translations of the phrase from both phrase tables and left the choice to the decoder.

We added three additional features to the new phrase table to avail the information about the origin of phrases (as in Nakov and Ng (2009)).

3.3 Evaluation and Discussion

We performed the following experiments:

- *S0* involved translating the EG' test using AR.
- *S1* and *S2* trained on the EG' with EG_{en} and both EG_{en} and *GW* for LM training respectively.
- S_* used phrase merging technique. All systems trained on both EG' and AR corpora. We built separate phrase tables from the two corpora and merged them. When merging, we preferred AR or EG' for S_{AR} and $S_{EG'}$ respectively. For S_{ALL}, we kept phrases from both phrase tables.

Table 2 summarizes results of using EG' and phrase table merging. *S0* was slightly better than *B1*, but lagged considerably behind training using EG or EG'. *S1*, which used only EG' for training showed an improvement of 1.67 BLEU points from the best baseline system (*B4*). Using both language models (*S2*) led to slight improvement. Phrase merging that preferred phrases learnt from EG' data over AR data performed the best with a BLEU score of 16.96.

We analyzed 100 test sentences that led to the greatest absolute change in BLEU score, whether positive or negative, between training with EG and EG'. The largest difference in BLEU was 0.69 in favor of EG'. Translating the Egyptian sentence "wbyHtrmwA AlnAs AltAnyp" وبيحترموا الناس التانية produced "وبيحترموا (OOV)

	Train	LM	BLEU	OOV
B4	EG	$EG_{en}GW$	14.23	5.2
S0	AR	EG'_{en}	8.61	2.0
S1	EG'	EG'_{en}	15.90	2.6
S2	EG'	$EG'_{en}GW$	16.10	2.6
S_{AR}	PT_{AR}	$EG'_{en}GW$	16.14	0.7
$S_{EG'}$	$PT_{EG'}$	$EG'_{en}GW$	**16.96**	0.7
S_{ALL}	$PT_{EG',AR}$	$EG'_{en}GW$	16.73	0.7

Table 2: Summary of results using different combinations of EG'/English and MSA/English training data

the second people" (BLEU = 0.31). Conversion changed "wbyHtrmwA" to "wyHtrmwA" and "AltAnyp" التانية to "AlvAnyp" الثانية, leading to "and they respect other people" (BLEU = 1). Training with EG' outperformed EG for 63 of the sentences. Conversion improved MT, because it reduced OOVs, enabled MADA+TOKAN to successfully analyze words, and reduced spelling mistakes.

In further analysis, we examined 1% of the sentences with the largest difference in BLEU score. Out of these, more than 70% were cases where the EG' model achieved a higher BLEU score. For each observed conversion error, we identified its linguistic character, i.e. whether it is lexical, syntactic, morphological or other. We found that in more than half of the cases (≈57%) using morphological information could have improved the conversion. Consider the following example, where (1) is the original EG sentence and its EG/EN translation, and (2) is the converted EG' sentence and its EG'/EN translation:

1. لان دي حسب رغبتك
   lAn dy Hsb rgbtk
   because this is according to your desire

2. لأن هذه حسب رغبته
   lOn h*h Hsb rgbth
   because this is according to his desire

In this case, "rgbtk" رغبتك ("your wish") was converted to "rgbth" رغبته ("his wish") leading to an unwanted change in the translation. This could be avoided, for instance, by running a morphological analyzer on the original and converted word, and making sure their morphological features (in this case, the person of the possessive) correspond. In a similar case, the phrase معنديش اعداء

(mEndy$ AEdA') was converted to عندي اعداء (Endy OEdA'), thereby changing the translation from "I don't have enemies" to "I have enemies". Here, again, a morphological analyzer could verify the retaining of negation after conversion.

In another sentence, "knty" كنتي ("you (fm.) were") was correctly converted to the MSA "knt" كنت, which is used for feminine and masculine forms. However, the induced ambiguity ended up hurting translation.

Aside from morphological mistakes, conversion often changed words completely. In one sentence, the word لبانه (lbAnh; "chewing gum") was wrongly converted to لأنه (lAnh; "because it"), resulting in a wrong translation. Perhaps a morphological analyzer, or just a part-of-speech tagger, could enforce (or probabilistically encourage) a match in parts of speech.

The conversion also faces some other challenges. Consider the following example:

- هوا احنا عملنا ايّيه
  hwA AHnA EmlnA Ayyyh
  he is we did we What ? ?

- هو نحن عملنا ايه
  hw nHn EmlnA Ayh
  he we did we do ? ?

While the first two words هوا احنا (hwA AHnA) were correctly converted to هو نحن (hw nHn), the final word ايّيه (Ayyyh; "what") was shortened but remained dialectal ايه (Ayh) rather than MSA ماذا/ما. There is a syntactic challenge in this sentence, where the Egyptian word order in interrogative sentences is normally different than the MSA word order, where the interrogative particle appears at the end of the sentence instead of at the beginning. Addressing this problem might have improved translation.

The above analysis suggests that incorporating deeper linguistic information in the conversion procedure could improve translation quality. It seems that using a morphological analyzer is one promising possibility. It should be possible to run a morphological analyzer on both the original EG sentence and the converted EG' sentence and then to compare the morphological features. Discrepancies should be probabilistically incorporated in the conversion. Exploring this approach is left for future work.

4 Conclusion

We presented an Egyptian to English MT system. In contrast to previous work, we used an automatic conversion method to map Egyptian close to MSA. The converted Egyptian EG' had fewer OOV words and spelling mistakes and improved language handling. The MT system built on the adapted parallel data showed an improvement of 1.87 BLEU points over our best baseline. Using phrase table merging that combined AR and EG' training data in a way that preferred adapted dialectal data yielded an extra 0.86 BLEU points. We will make the training data for our conversion system publicly available.

For future work, we want to expand our work to other dialects, while utilizing dialectal morphological analysis to improve conversion. Also, we believe that improving English language modeling to match the genre of the translated sentences can have significant positive impact on translation quality.

References

David Chiang, Mona T. Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for Arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, Maui, Hawaii, USA.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual Conference of the Association for Computational Linguistics*, Uppsala, Sweden.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, Seattle, Washington.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK.

Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed

Ammar. 2011. Improved transliteration mining using graph reinforcement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, Edmonton, Canada.

Emad Mohamed, Behrang Mohit, and Kemal Oflazer. 2012. Transforming standard Arabic to colloquial Arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Short Paper*, Jeju Island, Korea.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Short Paper*, Jeju Island, Korea.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio.

Wael Salloum and Nizar Habash. 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*, Edinburgh, Scotland.

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

Masao Utiyama and Hitoshi Isahara. 2008. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems*, Cairo University, Egypt.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, Portland, Oregon.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada.