Dial2MSA: A Tweets Corpus for Converting Dialectal Arabic to Modern Standard Arabic

Hamdy Mubarak
QCRI, Hamad Bin Khalifa University (HBKU), Doha, Qatar
hmubarak@hbku.edu.qa

Abstract

Modern Standard Arabic (MSA) is the official language used in formal communications while Dialectal Arabic (DA) refers to the spoken languages in different Arab countries and regions, and they are widely used on social media for daily communications. There are differences between DA and MSA at almost all levels, and resources for DA are very limited compared to MSA. In this paper, we present Dial2MSA corpus; the first and largest corpus of dialectal tweets with translations to MSA as provided by large number of native speakers through crowdsourcing. We describe how we collected the tweets, annotated them and measured translation quality. We aim that Dial2MSA can promote researches in understanding and quantifying differences between DA and MSA, dialect identification, converting DA to MSA (hence using MSA resources) and machine translation (MT) among other applications. Roughly, the corpus contains 5,500 and 5,000 tweets written in Egyptian and Maghrebi dialects with verified MSA translations (16,000 and 8,000 pairs in order), and 6,000 tweets written in Levantine and Gulf dialects with MSA translations (18,000 pairs for each without verification). The corpus is freely available for research purposes.

Keywords: Arabic Dialects, Dialect to MSA conversion, Parallel Corpus, Crowdsourcing

1. Introduction

Modern Standard Arabic (MSA) is the lingua franca of the Arab world, and it's used in official communications and speeches such as books, educational materials and newspapers. On the other hand, Dialectal Arabic (DA) refers to local dialects (or languages) spoken in different countries and regions, and they differ from country to another and sometime from city to another in vocabulary, morphology, and spelling among other things. These dialects are widely used on daily interactions and on social media platforms such as Facebook and Twitter.

Conventionally, researchers in the Arabic Natural Language Processing (NLP) field divide DA into major dialectal groups, namely: Egyptian (EGY), Maghrebi (MGR) spoken in the Maghreb region or North Africa, Levantine (LEV) spoken in the Levant, Gulf (GLF) spoken in the Arabic Peninsula, and Iraqi (IRQ). Sometimes IRQ is considers as one of the Gulf dialects.

There are many resources for MSA, such as large annotated corpora and tools, for different NLP tasks (e.g morphological analysis, parsing, machine translation, etc.) which generally achieve high scores. Compared to MSA, DA suffers from lack of resources. One possible solution for some tasks is to convert DA to MSA (i.e. use MSA as a pivot language or a bridge) such as researches done by (Bakr et al., 2008), (Al-Gaphari and Al-Yadoumi, 2010), (Sawaf, 2010), (Sajjad et al., 2013), (Salloum and Habash, 2013) and (Shaalan, 2016) to enhance translating DA to English.

Moreover, there is a lot of work in the MT field to convert from a resource-poor language to other languages by pivoting on a closely-related resource-rich language such as in (Durrani et al., 2010), (Hajič et al., 2000), and (Nakov and Tiedemann, 2012). This conversion can be done at different levels: character level transformation, word level translation or language-specific rules.

Dialect to MSA conversion or translation is usually performed using handcrafted rules and heuristics that require deep linguistic knowledge and extensive manual efforts. As reported by (Sajjad et al., 2013), conversion can also be done using translation methods but generally this requires parallel data (pairs of DA and MSA) which is not available. They manually created a lookup table of EGY-MSA words, and applied an automatic character-level transformation model to change EGY to something similar to MSA, and this gave a gain of 1.87 BLEU points for translating EGY to English.

In this paper, we introduce Dial2MSA; a new large-scale corpus of DA-MSA pairs of tweets for major dialects (EGY, MGR, LEV and GLF) as written by native speakers. We aim to support the field of dialectal NLP and reduce the effort of building linguistic rules for conversion by providing parallel data that can be used by statistical machine translation (SMT) techniques between these closely-related languages.

It is worth mentioning that Dial2MSA is different than the Arabic multi-dialectal parallel corpus published by (Bouamor et al., 2014) in different aspects:

- Bouamor's corpus contains translations of 2,000 EGY sentences to Palestinian, Syrian, Jordanian and Tunisian dialects in addition to MSA. Starting from EGY can be considered as biased input, and does not give the variety and naturalness found in native tweets written in these dialects.
- Each sentence in Bouamor's corpus is translated by only one person (the same person) per dialect, and in our corpus hundreds of native speakers participated in the translation process (multiple translations for each tweet) which guarantees wide range of opinions.

- Our corpus size is bigger.

Next sections have details about corpus collection, annotation and measuring translation quality. Then some statistics and examples are provided.

2. Data Collection

From a corpus of 175M Arabic tweets collected during March 2014[1], we filtered tweets using very strong dialectal words for each major dialect to extract dialectal tweets. These dialectal words (140 words) are mostly function words that are used exclusively in each dialect and they were revised by native speakers. Initial list was obtained from (Mubarak and Darwish, 2014b) then it was revised manually for better quality. Examples are shown in Table 1 and the full list can be downloaded from `http://alt.qcri.org/~hmubarak/EGY-MGR-LEV-GLF-StrongWords.zip`.

Dialect	Examples of dialectal words
EGY	ده، عاوز، إزاي this, want, how
MGR	بزاف، علاش، هكي very much, why, like this
LEV	هيك، مشان، عنجد like this, for, really
GLF	إشلون، شصار، مو بطبيعي how, what happened, not natural

Table 1: Strong dialectal words

For each dialect, we removed duplicate tweets, and selected tweets having lengths between 25 and 90 Arabic characters without counting mentions, URL's, etc. (roughly between 5 and 15 words), then selected random 6,000 tweets for the next annotation process.

3. Data Annotation

We created annotation jobs (Task1), one for each dialect, on CrowdFlower[2] (CF) where we showed dialectal tweets to annotators and asked them to provide corresponding MSA translations or conversions to have pairs of DA-MSA. Annotators were selected from the the countries that speak the target dialect (e.g. for MGR, annotators are restricted to be from Maghreb countries).

For quality control, we used the code and applicable best practices suggested by (Wray et al., 2015) and (Mubarak, 2017b) to prevent, as much as possible, bad annotations for different types of poor translation. Each dialectal tweet was converted to MSA by different annotators (5 for EGY and 3 for other dialects), and around 200 annotators contributed in each annotation task. This gives a wide diversity of opinions needed for such tasks. Figure 1 shows a sample EGY tweet and its MSA translations as provided by different annotators.

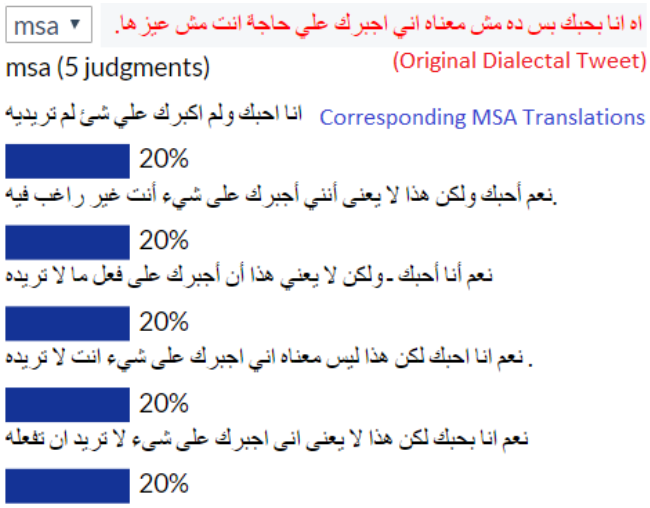

Figure 1: CF Task1: Converting DA to MSA

Quality of annotation at CF can be increased by using test questions where we provide their correct or gold answers, and annotators must pass a minimum threshold (typically 70%) of these test questions to continue. But because CF has limited capabilities in text comparison, and sentences can be expressed in many different ways, it's hard to list all possible forms of MSA sentences that can be used as gold answers to test questions. So to increase quality of the provided DA-MSA pairs, we created another annotation job for each dialect (Task2) to verify whether each pair is correct (i.e. having same meaning) or not. In this task, quality was controlled by using 50 test questions (correct pairs), and annotators should pass successfully a threshold of 80% to consider their work. Each pair was judged by 3 annotators who speak the target dialect. Sample annotation is shown in Figure 2.

Figure 2: CF Task2: Verify DA-MSA pairs

Translation jobs were completed for all dialects, and verification jobs of of the collected pairs were launched and completed for EGY and MGR because there are many annotators from theses regions (33% and 30% in order as obtained from recent surveys for Arab annotators on CF (Mubarak and Darwish, 2016)). We plan to verify collected pairs for other dialects as well.

[1]Using Twitter API (`http://dev.twitter.com`) with language filter assigned to "lang:ar"

[2]Crowdsourcing platform: `www.crowdflower.com`

Figures 3, 4, 5, and 6 show examples of dialectal tweets for each dialect and their MSA translations as provided by annotators. Dialectal words and their equivalent MSA words are marked with different colors.

EGY to MSA	EGY
ليس هناك احسن من الافراد الذين يحفظون السر	مفيش احلى من الناس اللى بتحفظ السر
ليس هناك اجمل من الاشخاص الذين يحفظون السر	مفيش احلى من الناس اللى بتحفظ السر
لا يوجد أحلى من الناس التي تحتفظ بالسر	مفيش احلى من الناس اللى بتحفظ السر

Figure 3: Example of EGY to MSA conversion

MGR to MSA	MGR
أنا لم أعد اتحمل أكثر من هذا	انا معش حا نتحمل اكثر من هكي
انا لن اتحمل اكثر من هكذا	انا معش حا نتحمل اكثر من هكي
انا لناتحمل اكثر من هذا لا استطيع	انا معش حا نتحمل اكثر من هكي

Figure 4: Example of MGR to MSA conversion

LEV to MSA	LEV
الهي لا تحرمني من هكذا اصدقاء	الهي ما انحرم من هيك صحاب
اللهم لا تحرمني من هكذا أصحاب	الهي ما انحرم من هيك صحاب
يا رب لا تحرمني من هؤلاء الاصدقاء	الهي ما انحرم من هيك صحاب

Figure 5: Example of LEV to MSA conversion

4. Data Quality

To get a rough estimate about the quality of obtained translations, we randomly selected 100 EGY tweets and their verified MSA translations (410 sentences), and asked a professional linguist to do needed corrections to make MSA sentences free of spelling and grammar errors, and retain the whole meaning of original tweets[3].

For comparison, we normalized MSA translations, before and after linguistic revision, to solve common spelling mistakes in some letters. For example, we converted all shapes of Hamza to plain Alif, Alif Maqsoura to dotted Yaa, and Taa Marbouta to Haa (الأخطاء الشائعة في الهمزات والتاء المربوطة والألف المقصورة), and removed punctuation marks. Correcting such errors is fairly easy and can achieve high accuracy by consulting a large clean corpus such as of `Aljazeera.net` as shown in (Mubarak and Darwish, 2014a). The overlap between translations before and after linguistic revision was 90% indicating high annotation quality obtained from non-experts.

[3]Linguistic corrections can be downloaded from: `http://alt.qcri.org/~hmubarak/EGY2MSA-sample-correction.zip`

GLF to MSA	GLF
الذي يريدك يعرف كيف يحافظ عليك	إلي يبيك...يعرف اشلون يحافظ عليك
الشخص الذي يريدك يعرف جيدا كيف يحافظ عليك	إلي يبيك...يعرف اشلون يحافظ عليك
من يريدك يعرف كيف يحافظ عليك	إلي يبيك...يعرف اشلون يحافظ عليك

Figure 6: Example of GLF to MSA conversion

Figure 7 shows examples of MSA translations obtained from CF, and their corrections for the EGY tweet:
احساس حلو اوي لما تلاقي حد بيقولك أنا بفرح اوي لما بتكلم معاك.
Spelling and grammar errors and their corrections are marked in different colors. Most errors are common and can be recovered, and there are some grammatical errors (case ending) and few split/merge errors. We estimate MSA translations for other dialects to have similar accuracy and they all need spelling correction.

We noticed that some translations are a bit unnatural, and this can be checked probably by using language models trained on MSA. We leave this for future work.

For tweets having multiple translations, if we want to get the best translation with minimum errors, we can use ROVER algorithm to combine these translations. ROVER (Recognizer output voting error reduction) (Fiscus, 1997) is used in automatic speech recognition to implement a "voting" or rescoring process for combining outputs of multiple speech recognizers (translations in our case). It seeks to reduce word error rates by exploiting differences in the nature of the errors in multiple outputs as shown in Figure 8.

5. Preliminary Data Analysis

Statistics about Dial2MSA corpus are listed in Table 2 and it can be downloaded from `http://alt.qcri.org/~hmubarak/EGY-MGR-LEV-GLF-2-MSA.zip`.

We started by 6,000 tweets for each dialect, and approximately for EGY and MGR, we obtained 5,500 and 5,000 tweets[4] with 16,000 and 8,000 verified MSA translations respectively, i.e. almost half the annotations of Task1 were approved in Task2. For LEV and GLF, we have 18,000 MSA translations per each and they need verification.

For the verified DA-MSA pairs for EGY and MGR, we calculated number of words, average number of words per sentence, and the Overlap Coefficient (OC) (#common words in DA and MSA / minimum length) as suggested by (Bouamor et al., 2014) for normalized words. Results are shown in Table 3. Their OC values for Egyptian and Tunisian dialects are 0.45 and 0.31 in order.

6. Resource Description and Benefits

In this paper, we created Dial2MSA; a corpus of parallel pairs of DA tweets and their conversions or translations to

[4] All translations of some tweets were rejected

EGY to MSA (Linguistic Revision)	EGY to MSA (CrowdFlower)
إحساس جميل جدا حين تجد من يقول لك : أنا سعيد جدا حين أتحدث إليك	احساس جميل جدا حين تجد من يقول لك انا سعيد جدا حين أتحدث اليك
إحساس جميل حينما أجد أحدًا يقول لي : أنا سعيد حينما أتحدث معك	احساس جميل حينما اجد حد يقول لي انا سعيد حينما اتحدث معك
إحساس جميل عندما : شخص يخبرك أنه يسعد جدا عندما تتحدث معه	احساس جميل عندما شخص يخبرك ان يسعد جدا عندما تتحدث معه
إنه إحساس جميل عندما تجد شخصًا يخبرك بأنه فرح عند التحدث معك	انه احساس جميل عندما تجد شخص يخبرك بانه فرح عند التحدث معك
شعور جميل جدا أن تجد من يقول لك : أنا أفرح جدا بحديثي معك	شعورجميل جدا ان تجد من يقول لك انا افرح جدا بحديثى معك

Figure 7: Linguistic revision example of MSA translations

Dialect	#Orginial Tweets	#MSA (Task1)	#Verified MSA (Task2)	%	#Tweets having MSA	Average #MSA/Tweet
EGY	6,000	30,000	16,355	55%	5,565	2.94
MGR	6,000	18,000	7,912	44%	4,953	1.6
LEV	6,000	18,000	-	-	-	-
GLF	6,000	18,000	-	-	-	-

Table 2: Statistics about Dial2MSA corpus

there's	a	lot	of	@	like	societies	@	@	ruin	engineers	and	lakes
there's	the	labs	@	@	like	societies	@	for	women	engineers	i	think
there's	the	last	@	@	like	societies	@	true	of	engineers	and	like
was	@	alive	@	the	legal	society	is	for	women	engineers	and	like
there's	a	lot	of	@	like	society's	@	@	through	engineers	@	like

REF: there's a lot OF like societies for women engineers and like
HYP: there's a lot ** like societies for women engineers and like

Figure 8: Aligning different outputs using ROVER

MSA as obtained from native speakers. We used crowdsourcing platform with quality control settings applied at different levels to have high quality of annotations with a wide variety of opinions which is normally not available in traditional companies. The cost of annotation jobs is less expensive and progress is fast compared to normal workers, and quality is comparable to language experts.

The obtained parallel DA-MSA pairs can help in understanding and quantifying similarities and differences between DA and MSA at different levels (phonology, morphology, and syntax), and enhancing dialectal Arabic NLP. Conversion was applied at sentence level (i.e. context is considered) which gives high accuracy.

Mapping between DA and MSA at different levels (characters, words or patterns) can be obtained automatically with high accuracy using alignment techniques because in most cases, there are no much differences in word order between them. This reduces the need for writing linguistic rules for DA to MSA conversion which requires a lot of experience and effort. For example, we can use Smith-Waterman algorithm[5] to align dialectal words and MSA counterparts with high accuracy as shown in Figure 9.

MSA words can also be used as pivots to align dialectal words in different dialects having the same meaning, ex: نحن = إحنا، نحنا، حنا.. (writing variations of "we" in MSA and DA).

EGY	MSA	MGR	MSA
معظم	معظم	علي	على
النصايح	النصائح	فكرا	فكرة
اللى	التي	تتنرفزو	هل
بننصح	ننصح		تغضبون
بيها	بها	مني	مني
غيرنا	غيرنا	لما	عندما
احنا	نحن	نسال	اسأل
مبنعملهاش	لا	هذي	هذه
	نفعلها	شن	ما
			هو
		معناها	معناها
		وهكي	وهكذا

Figure 9: Examples of aligning DA and MSA

7. Conclusion and Future Work

In this paper, we presented Dial2MSA; a corpus of DA tweets and their translations to MSA. This is the first and largest corpus available for DA to MSA conversion where original raw tweets are written by native speakers for each dialect which gives the needed naturalness and diversity found on social media sites.

[5] https://en.wikipedia.org/wiki/Smith-Waterman_algorithm

Dialect	#Words (Tweets)	#Words (MSA)	#Unique Words (Tweets)	#Unique Words (MSA)	#Words/sentence (Tweets)	#Words/sentence (MSA)	**Overlap Coeff.**
EGY	77,800	206,989	17,399	31,288	13.9	12.6	**0.33**
MGR	53,351	85,557	18,856	19,908	10.7	10.8	**0.38**

Table 3: Statistics about verified DA-MSA translation pairs

Translations of tweets are provided by native speakers through crowdsourcing, and each tweet is translated (and translations are verified) by different annotators to have variety of opinions. We measured quality of samples from the obtained pairs and showed that it's comparable to quality of language experts. The corpus is freely available for research purposes.

We plan to study the usefulness of this corpus on automatic translation of DA to MSA, translation across dialects, and from DA to English through pivoting on MSA. Also, we plan to correct spelling and grammar mistakes in the annotations and revise the automatic alignment to have more accurate and rich data.

It's worth mentioning that in (Mubarak, 2017a), translating EGY to MSA was applied at word level (i.e. lookup table) without having translations of complete tweets. For example, the word بس was translated to فقط، لكن (only, but). We estimate that translating complete tweets (such as in Dial2MSA corpus) would be more useful, and can produce a more fluent translation to MSA, and therefore better translation to English for example. Besides, using alignment algorithms can extract entries in DA-MSA lookup tables accurately especially for common words. Benefits of using translations of complete tweets over (or maybe with) individual words need to be experimented.

8. Bibliographical References

Al-Gaphari, G. and Al-Yadoumi, M. (2010). A method to convert sana'ani accent to modern standard arabic. *International Journal of Information Science & Management*, 8(1).

Bakr, H. A., 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*.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

Durrani, N., Sajjad, H., Fraser, A., and Schmid, H. (2010). Hindi-to-urdu machine translation through transliteration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 465–474. Association for Computational Linguistics.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE.

Hajič, J., Hric, J., and Kuboň, V. (2000). Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, pages 7–12. Association for Computational Linguistics.

Mubarak, H. and Darwish, K. (2014a). Automatic correction of arabic text: a cascaded approach. *ANLP 2014*, page 132.

Mubarak, H. and Darwish, K. (2014b). Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Mubarak, H. and Darwish, K. (2016). Demographic surveys of arab annotators on crowdflower. In *Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms Workshop (WebSci16)*.

Mubarak, H. (2017a). Analysis and quantitative study of egyptian dialect on twitter. In *The 3rd International Workshop on Natural Language Processing for Informal Text (NLPIT 2017), Maastricht, Italy*.

Mubarak, H. (2017b). Crowdsourcing speech and language data for resource-poor languages. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. AISI 2017, Advances in Intelligent Systems and Computing, vol 639. Springer, Cham*.

Nakov, P. and Tiedemann, J. (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 Papers-Volume 2*, pages 301–305. Association for Computational Linguistics.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating dialectal arabic to english. In *ACL (2)*, pages 1–6.

Salloum, W. and Habash, N. (2013). Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *HLT-NAACL*, pages 348–358.

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*.

Shaalan, K., B. H. Z. I. (2016). Transferring egyptian colloquial dialect into modern standard arabic. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2007), John Benjamins 2016*.

Wray, S., Mubarak, H., and Ali, A. (2015). Best practices for crowdsourcing dialectal arabic speech transcription. In *ANLP Workshop 2015*, page 99.