Twitter in the Context of the Arab Spring

Muhammad Abdul-Mageed
Department of Linguistics & School of Informatics and Computing
Indiana University, Bloomington
mabdulma@indiana.edu

Christopher Brown
Department of Linguistics
University of Texas at Austin
chrisbrown@utexas.edu

Dua'a Abu-Elhij'a Mahajna
Department of Linguistics
Indiana University, Bloomington
dabuelhi@umail.iu.edu

Abstract
Except for some initial analyses (Abdul-Mageed & Albogmi, 2011; Abdul-Mageed, Albogmi, Gerrio, Hamed & Aldibasi, 2011; Abdul-Mageed & Abu Mostafa, 2012; Abdul-Mageed, 2013), Arabic Twitter has not been the target of much research. Although these initial analyses have accumulated insights as to some Twitter user practices and the micro-blogging service's content, these studies remain limited in terms of scale (i.e., they did not depend on huge datasets as is attempted in the current work), nor did they examine Twitter use and practices in the context of the Arab Spring. In the current work, our goal is to bridge these gaps. In other words, we investigate (1) the employment of various Arabic varieties, (2) sentiment expression, and (3) topic distribution in two Twitter sub-corpora, one pre-dating the Arab Spring and another occurring in its context. Our analyses show the impact of the political events on the social network at the three fronts.

Keywords
Arab Spring; Twitter; computer-mediated communication; sentiment; Arabic; Arabic dialects; Modern Standard Arabic; social media; internet activism.

Introduction

Twitter, the micro-blogging service, has become popular in many countries as a communication platform for posting updates about personal activities, expressing opinions, sharing experiences, broadcasting news, etc. However, although there is currently a flurry of interest in researching the English-language Twitter, as yet not much research has been done on other languages used on the micro-blogging service, and Arabic is no exception. In the current paper, we report attempts to partially bridge this gap by investigating the use of Arabic Twitter both before *and* in the context of the Arab Spring. Motivations for studying Arabic Twitter include (1) the popularity of Twitter as a networking and communication tool during the ongoing Arab Spring, (2) that the language is very widely spoken (with about 300 million native speakers in 22 countries), and (3) the fast growing rate of Arabic on the Web (with Arabic rated as the world's fastest growing language on the Web in 2009 with 2,297.7% growth rate, http://www.internetworldstats.com/stats7.htm).

Goal and Research Questions

Our goal in the current work is to analyze the linguistic features, identify language variety employed (i.e., Modern Standard Arabic [MSA], the modern standard variety of the language, vs. dialects [e.g., Egyptian, Levantine, Moroccan]), topics, and sentiments expressed in Arabic tweets. We thus aim to answer the following specific research questions:

RQ1: What are the linguistic features of Arabic as used in Twitter?
RQ2: What is the distribution of Modern Standard Arabic vs. dialects (i.e., all dialects as a single category) in Arabic Twitter?
RQ3: What is the distribution of Arabic dialects in Arabic Twitter?
RQ4: What are the topics that Arabic users tweet about?
RQ5: To what extent are Arabic tweets opinionated and what is the distribution of sentiment (e.g., positive, negative) in opinionated tweets?

Data Collection, Sampling, and Methods

We employ corpus linguistics (McEnery & Wilson, 2001) and content analysis (Bauer, 2000) methods on two Arabic Twitter corpora, one collected before the Arab Spring and one after it. The first sub-corpus is composed of 233,309 tweets (henceforth the Twitter Arabic Corpus [TAC]) automatically extracted from a two-month Twitter stream between November 11th, 2009 and February 1st, 2010 as reported in Petrovic, Osborne & Lavrenko (2010). The second sub-corpus is composed of about 50,000,000 tweets (henceforth the Twitter Arabic Spring Corpus [TASC]) collected between February 1st, 2012 and March 1st, 2013. For the manual content analysis, we employ a grounded-theory approach to analyze a random sample of 2,000 tweets from each sub-corpus (making up a total of 4,000 tweets).

Results

Results of the analysis of linguistic features show that Arabic tweets do not differ strikingly from 'offline' Arabic, although they share some features characteristic of some varieties of online language (e.g., paralinguistic and prosodic features). Regarding the analysis of language variety, it was found that 66.1% of the tweets were in MSA, whereas the remaining 33.9% were in Arabic dialects. Initial nuanced analyses of dialects show that dialects associated with countries where revolutions/protests have taken place (e.g., Egyptian Arabic) are represented with higher rates than other dialects.

The sentiment analysis indicates that 35.44% of the tweets are solely factual and bear no sentiment. Out of the 65.66% non-factual/opinionated tweets, 25.28% are positive, 17.22% are negative, 9.40% are neutral, and 2.78% are mixed. Initial analyses of TASC show more negative than positive content. An analysis of the topics of tweets in the TAC sub-corpus revealed that 17.25% of users tweeted about cultural issues, 14.80% about politics, 7.40% about feelings & activities, and the rest about other miscellaneous (e.g., religious, economic, and educational) topics. Initial analyses of topics in TASC quite predictably show a significant increase of especially political content.

Discussion and Conclusion

At first glance, the finding that the 140-character limit imposed by Twitter does not force users to frequently employ specific linguistic features (e.g., clippings and abbreviations) is surprising. However, this is expectable as in Arabic the practice of dropping vowels results in a level of ambiguity and employing clippings and abbreviations would heighten such ambiguity and render text unintelligible. We thus note that social media Arabic is different from digital forms of other Indo-European languages (e.g., English), not only because of cultural differences but also because of the linguistic features of the language itself (e.g., its orthographic underspecification). The finding that the majority of tweets are in MSA is equally surprising. This specific language variety distribution is perhaps due to the fact that several news organizations use Twitter to post headlines of their news stories. In addition, many news anchors and highly educated individuals are among the Arabic Twitter users. The finding that more positive than negative tweets are posted shows that Twitter is different from e.g., listserves (Thompsen & Ahn, 1992) where users are less likely to know one another. The more negative use of the TASC is indicative of the employment of the Twitter medium for argumentative causes (e.g., causes characteristic of the Arab Spring). The fact that cultural and political tweets are the most frequent in TAC indicates that Arabic Twitter users are less interested in personal issues than they are in community-based issues. This claim can perhaps be accounted for by the ongoing revolutions in the Arab world in which Arabic Twitter is playing a major role. Perhaps the period from which TAC was collected can be viewed as setting the stage for the revolutions. The more political content and the heightened proportion of dialects representing regions with political changes in TASC are intuitive yet illustrative results, which show the role that social media in general, and Twitter in particular, has been playing in the Arab Spring.

Sources

Bauer, M. (2000). Classical content analysis. In M. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 131-151). Thousand Oaks, CA: Sage.

Abdul-Mageed, M., Albogmi, H. (2011). Taghreed?: What Arabs say on Twitter and how they say it. Paper presented at the Georgetown University Round Table on Languages and Linguistics (GURT2011): Language and New Media: Discourse 2.0, March 9-12, Washington DC, USA.

Abdul-Mageed, M., Albogmi, H., Gerrio, A., Hamed, E., & Aldibasi, O. (2011). Tweeting in Arabic: What, how and whither. Paper presented at the 12th annual conference of the Association of Internet Researchers (Internet Research 12.0 – Performance and Participation), October 10-13, Washington, Seattle, USA.

Abdul-Mageed, M., & Abu Mostafa, H. (2012). Linguistic features, language variety, and sentiment in online Arabic. *Pragmatics Festival*, April 19-21, Indiana University, Bloomington, USA.

Abdul-Mageed, M. (2013). Social media Arabic. Paper presented at the 27th Annual Symposium on Arabic Linguistics, Feb. 28 - March 2, Indiana University, Bloomington, IN, USA.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (Vol. 1). Edinburgh.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010, June). The Edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media* (pp. 25-26).

Thompsen, P.A., & Ahn, D.K. (1992). 'To Be or Not To Be: An Exploration of E-Prime, Copula Deletion and Flaming in Electronic Mail.', Et Cetera: A Review of General Semantics, 49, 146–164.