Language Processing for Arabic Microblog Retrieval

Kareem Darwish, Walid Magdy, and Ahmed Mourad
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
{kdarwish, wmagdy, amourad}@qf.org.qa

ABSTRACT
The use of social media has profoundly affected social and political dynamics in the Arab world. In this paper, we explore the Arabic microblogs retrieval. We illustrate some of the challenges associated with Arabic microblog retrieval, which mainly stem from the use of different Arabic dialects that vary in lexical selection, morphology, and phonetics and lack orthographic and spelling conventions. We present some of the required processing for effective retrieval such as improved letter normalization, elongated word handling, stopword removal, and stemming.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms
Algorithms, Experimentation, Languages.

Keywords
Arabic Retrieval; Microblog Search; Arabic Twitter; Dialect Arabic Normalization

1. INTRODUCTION
Social media has been instrumental in facilitating the launch of the so-called "Arab Spring". The penetration of social media has been steadily increasing. The number of Facebook users in the Arab countries is estimated to be 42.4 million, representing 14.8% of the population. This number has increased by 10% between September 2011 and February 2012[1]. We estimate the number of Arabic microblogs, a.k.a. tweets, on Twitter to be in excess of 2 million tweets per day (based on continuous querying on Twitter).

With the increasing adoption of social media, there is a growing need for effective retrieval of relevant social content. This need has prompted the creation of the Microblog track in the 2011 Text REtrieval Conference (TREC). The track has focused exclusively on retrieving English tweets. The track's main task involved finding up to 30 relevant tweets per topic for 50 topics from roughly 5 million English tweets [12].

In this paper, we focus on the required processing for effective retrieval of Arabic microblogs, specifically Arabic tweets from Twitter. We use "tweet" and "microblog" interchangeably in the paper. Arabic tweets show many interesting linguistic phenomena that warrant special handling. These phenomena include:

1. The use of dialectic text as opposed to Modern Standard Arabic (MSA), which complicates retrieval because they lack spelling standards, causing a word to have multiple spellings. Also, dialects introduce a variety of new stopwords, and different dialects may make different lexical choices for concepts – often with transliterated words from other languages.
2. Microblog authors often employ word elongations (using repeated letters), word compressions (by omitting letters), and word decorations using non-Arabic letters.

In this paper, we explore the effect of some of these phenomena on information retrieval in the context of a collection of 112 million Arabic tweets that we collected from Twitter with an associated set of 35 topics and their relevance judgments. We test the effectiveness of state-of-the-art Arabic retrieval techniques, which are mostly geared for MSA, including orthographic and morphological processing. We introduce orthographic processing that is better suited for Arabic microblog text. Such processing includes improved character normalization, elongated and shortened word handling, expanded stopword removal with stopwords for different Arabic dialects, and stemming.

The contribution of this paper is as follows:

1. To the best of our knowledge, this is this first study on Arabic microblog retrieval.
2. We describe novel character normalization and word elongation and compression handling.
3. We explore the effect of stemming on Arabic tweet retrieval, and we provide a stopwords list that covers different dialects.

2. BACKGROUND
2.1 Dialects in Arabic Microblogs
Though most Arabic speakers can read and understand MSA, they generally use different Arabic dialects in their daily interactions. With the spread of online social interaction in the Arab world, these dialects started finding their way to written online social interaction. There are 6 dominant dialects, namely Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni. Aside from those, there are many more Arabic dialects along with variations within them [17]. The pronunciations of different letters are often different in different dialect. One of the letters with most variation in pronunciation is "ق (q[2])". In MSA, it is typically pronounced similar to the English "q" as in "quote". However, it is pronounced as an "a" in Egyptian and Levantine (as in the "Alpine"), as a "ga" in Gulf (as in "gavel"), and as a "gh" in Sudanese (as in "Ghana"). Phonetic variations sometimes affect the way people spell words in microblogs. For example, the word "نظافة" (nZAfp) meaning "cleanliness" is often written as "نضافة" (nDAfp) to match the Egyptian dialect. Different dialects introduce new function words (usually stopwords) that don't exist in MSA. For example, the MSA word "هكذا" (hk*A) – meaning

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

[1] http://www.socialbakers.com
[2] Buckwalter transliteration is used exclusively in the paper.

"like this" – is expressed as "كده" (kdh), "شذي" ($*y), and "هيك" (hyk) in the Egyptian, Gulf, and Levantine dialects respectively.

In what follows are other noteworthy aspects that we do not handle in this paper. One such aspect affects lexical choices of dialects and the frequent borrowing from English and French. Different dialects make different lexical choices to express the same concepts. For example, "I want" is expressed as "عاوز (EAwz)" in Egyptian, "أبغي (|bgy)" in Gulf, and "بدي (bdy)" in Levantine. The most popular MSA form is "أريد (|ryd)".

Different dialects introduce new morphological patterns that don't exist in MSA. For example, Egyptian and Levantine Arabic use a negation construct similar to the French "ne-pas" negation construct. The word "ملعبتش (mlEbt$)", meaning "did not play", is composed of "m+ lEbt +$". Such morphological constructs almost exclusively affect verbs. A more elaborate treatment of Arabic dialects and their processing can be found in [5, 17].

2.2 Orthography in Arabic Microblogs

There are several orthographic features that are found in Arabic microblogs, including:

1. The frequent use of word elongations using repeated letters such as "مبروووك (mbrwwwk)" – meaning congratulations.
2. The ubiquity of spelling mistakes.
3. The use of words that express emotions such as laughter (ex. "لول (lwl)" corresponding to LOL in English).
4. The use of similar looking letters from other languages, such Farsi, to decorate words. Consider the word "ﻳﺮﻗﺪوا" which is meant to represent "يرقدوا (yrqdw)" – meaning "they settle".

All these issues, some of which are not unique to Arabic, complicate retrieval and we attempt to handle them in this paper.

2.3 Arabic Retrieval

Most studies are based on a single large collection from the TREC-2001/2002 cross-language retrieval track [4, 10]. The studies examined indexing using words, word clusters [7], morphological analysis [4, 7], and character n-grams [4]. The effects of normalizing alternative characters and removing diacritics and stopwords have also been explored [4]. The studies suggest that light stemming, character n-grams, and statistical stemming are the best index terms. Since Arabic morphology is ambiguous, stemming attempts to remove the most likely clitics (prefixes and suffixes) and to conflate inflected word forms and some derivational forms. An Arabic stemmer of 97.1% accuracy was shown by Darwish et al. [1] to lead to statistical improvements over using light stemming. Diab [2] used an SVM classifier, trained on the Arabic Penn Treebank, to ascertain the optimal segmentation for a word in context with an accuracy of 99.2%. Although consistency is more important for IR applications than linguistic correctness, improved correctness would naturally yield to greater consistency. In this paper, we used a system akin to Diab [2] as a baseline. We are not aware of any previous work on Arabic social search or microblog search.

2.4 Microblog Retrieval

Interest in microblog retrieval has increased in recent years, with microblogs being used in a variety of tasks. For example, [15] used tweets as a news source and compared them to other online news media. Phelan et al. [13] used tweets to recommend news to users. Naveed et al. [9] illustrated the challenges of microblog retrieval, where documents are very short and typically focus on a single topic. Taveen et al. [16] compared microblog queries to web queries. They found microblog queries usually represent a user's interest to find updates about a given event or personality as opposed to finding relevant pages about a topic.

Recently, TREC introduced a Microblog track focused on English microblog retrieval [12]. The track provided a collection of about 14 million tweets with a set of 50 topics and their relevance judgments. The task was to search the English subset of the collection (roughly 5 million tweets) for relevant tweets. The evaluation metric used for evaluation was precision at 30 (P@30), which was picked based on the assumption that users usually checks no more than 30 tweets per query. Generally, simpler retrieval approaches performed at par or better than more sophisticated ones [12]. The best approaches included using pseudo relevance feedback and simple ranking features, such as term frequency, inverse document frequency, and tweet length [8, 11]. Other methods that used natural language processing did not significantly impact results [18]. Wei et al., [18] used: word compression for removing three or more repetitions of a character in a word; slang word translation, which replaced dialectic terms with proper English ones using a slang dictionary; spell checking using aspell; and replacement of OOV terms with similar English words, where similarity was based on character and phonetic edit distance, and contextual features. Although previous work showed that such processing was effective in correcting such erroneous words in tweets [6], [18] did not report significant improvements in retrieval effectiveness due to such processing. In this paper, we explore the use of similar linguistic processing for Arabic tweets.

3. TWITTER ARABIC COLLECTION

The collection contains a little over 112 million Arabic tweets that were scraped from Twitter between Nov. 20, 2011 and Jan. 9, 2012. We collected the tweets by issuing the query "lang:ar" against Twitter. For topics, we randomly picked 3,000 tweets to guide query construction. They gave a general indication of the topics in the collection. We formulated 35 topics, with each topic having a title query and a relevant exemplar. The queries ranged in length between 2 and 5 words with an average query length of 3.2 words. Consider the example query:

Title query: حرية استخدام الإنترنت (*Internet Freedom*)
Relevant exemplar:
Author ID: free_moment; Tweet ID: 145401039397453824;
Date: Sat Dec 10 10:14:43 AST 2011;
Text: وكالـة معا الاخباريـة: الاتحاد الأوروبي: خطوط عريضة لحرية استخدام الانترنت ستحدد قريباً - #Syria #EU http://t.co/RdHGNf8B
Translation: *(Ma'an News Agency: European Union: General guidelines for internet freedom will be announced soon)*

For relevance judgments, we manually judged all top 30 results for all our runs. Duplicate or near duplicate results were eliminated, and judgments were propagated from judged tweets to all duplicate or near duplicate tweets. We deemed two tweets to be duplicates (or near duplicates) if the edit distance between the two tweets was less than 10% of the length of the shorter of the two tweets. Prior to computing the edit distance, we tokenized the tweets using our custom tokenizer, and we removed URL's, name mentions, hashtags, punctuations, emoticons, and symbols indicating laughter (such as lol). We made binary judgments of either 1 or 0 for 19,824 query-document pairs (566 per query on average), with an average of 267 relevant tweets per topic.

4. ARABIC PROCESSING

4.1 Basic Arabic Normalization

The Arabic alphabet consists of 28 letters. However, 8 additional characters are used in the Arabic text, and they

represent variant forms of some of the letters in the Alphabet or common replacements. These include:

- Different forms of *alef*: "آ (>)", "أ (|)", "إ (<)", and "ا (A)". Though their use depends on the morphology and context of the word, they are often erroneously used interchangeably.
- *alef maksoura* "ى (Y)", which is a form of *alef*, but is often confused in writing with the letter *ya* "ي (y)".
- *ta marbota* "ة (p)" which is often confused with *ha* "ه (h)".
- Different forms of *hamza*: " ئ (})", "ؤ (&)", and "ء (')", are used interchangeably depending on the role of a word in the sentence. Consider "سماؤه (smA&h)", "سماءه (smA'h)", and "سمائه (smA}h)" – meaning "his sky" – where "sky" is used as subject, object, or idafa (possessive construct) respectively.

Hence, the following normalizations are effective:

Alef: {أ, آ, إ} → ا; *Hamza*: {ئ, ؤ} → ء; *Alef maksoura*: {ى} → ي; *Ta Marbota*: {ة} → ه

Additionally, there are characters that require removal, namely:

- Kashida which is used for decorative word elongation.
- Optional diacritics, which represent short vowels in Arabic.

Except for the normalization of the different *hamza* forms, such processing was shown to be effective in the literature. The normalization *hamza* was shown to be effective in unpublished work and is commonly used in web search engines. We will henceforth refer to this processing as "basic normalization".

4.2 Extended Arabic Normalization

Basic normalization may not be sufficient for improved retrieval effectiveness for Arabic microblogs. An extended normalization steps are proposed to achieve better effectiveness.

4.2.1 Handling non-Arabic characters

Unlike MSA, Arabic microblog text sometimes contains non-Arabic characters for decoration ("پرقدوآ"). These additional characters are typically borrowed from Farsi and Urdu. We examined 10% of the tweets in the collection, and we made a list of all unique characters in the sub collection. We assigned mapped characters that were neither Arabic nor punctuation to Arabic characters. The top 10 such characters and their mappings are in Table 1. The full list of unique characters contained more than 2,000 non-Arabic characters, with half of them appearing less than 10 times. Only 385 of them could be mapped to Arabic characters. The remaining characters were mapped to null as they represented decorative diacritics and unrecognized characters.

Table 1: Frequent non-Arabic characters and their mappings

Character	گ	ھ	ہ	۰	أ	ا	ۆ	ـ	آ	ڵ
Mapped to	ك	ه	ه	0	/	1	و		/	ل

4.2.2 Handling elongated and shortened words

In many languages, words in microblogs are routinely elongated by repeating some of the characters in the word to express emotions or importance. For example, you may find words such as "cooooooool" and "loooool" in English tweets. In Arabic microblogs, we observed two related phenomena in tweets:

1. Some letters are often repeated multiple times.
2. Some repeated letters in valid words are routinely omitted. Example: "سعودين (sEwdyn)" was shortened from "سعوديين (sEwdyyn)" (meaning Saudis).

We developed an algorithm to handle these two cases. Given a word in the tweet *W*, a compressed form of it *C* is generated by omitting any repetition in characters. A compressed form C_i can represent a class that contains all surface forms of words W_j that leads to the same compressed form C_i. For example, the words: {cool, cooooool, coool, coollllll, col}, will all have the compressed form "col", that will act as a class including them all. All words in the tweet collection are classified according to their compressed form, and the most frequent surface form with each compressed form is saved in a hash table. Handling elongations and compressions for a word W_i works as follows:

1. If word W_i exists in an MSA dictionary, then it is left as it is, even repeated characters exists.
2. Else, a compressed version of the word is generated (even if the word has no repeated characters, i.e. $C_i = W_i$). If it exists in the table of compressed forms, then it is replaced with the most frequent surface form, which can contain more letter repetitions than original word W_i.
3. If compressed form C_i doesn't exist in hash table, then word W_i is directly replaced by the compressed form C_i.

In our experiments, we constructed the MSA dictionary from a collection of 349k articles from Aljazeera.net news site. We used articles from Aljazeera.net, because they rarely have misspellings.

Table 2: Different incarnations of the word "الله (Allh)"

Term	*Frequency*	*(translation)*	*MSA*
الله	11,304,952	*beautiful*	Yes
اله	645,560	*God*	Yes
اللله	10,541	*beautiful+*	No
االله	4,501	*beautiful+*	No

Table 2 shows an example of a set of words that have the same compressed form. As shown in Table 2, the first two terms are correct MSA words. Hence, they will be kept as is. Other terms are invalid terms. Therefore, they will be replaced with the most frequent term in the table whenever seen in a tweet or a query. Below is an example before and after applying the algorithm:

Before: يااا اللللله ايششش اللي السعودين يسوووه ده هههه
After: يا الله ايش اللي السعوديين يسوه ده هههههه
Translation: *Oh my God, what are the Saudis doing hahahaha*

5. EXPERIMENTATION

5.1 Experimental Setup

All experiments were performed on the aforementioned collection of 112 million tweets. We performed 7 runs as follows:

- **Baseline runs**: ***Norm*****:** only basic normalization is performed. ***Stem*****:** basic normalization and statistical stemming are performed. ***Char4g*****:** character 4-grams are used instead of words with basic normalization.
- **New runs**: ***NormD*****:** extended normalization is performed. ***StemD*****:** extended normalization and statistical stemming are performed. ***NormDS*****:** same as NormD with stopwords removed. ***StemDS*****:** same as StemD with stopwords removed.

Statistical stemming was performed using a reimplementation of the tokenizer in [2]. We constructed a stopword list that combines both 162 MSA stopwords that we acquired from the University of Neuchâtel[3] and 90 dialectic stopwords. To identify the dialectic stopwords, all the words in the tweet collection were sorted according to their document frequency and we manually examined the top 200 words to determine if they were indeed stopwords or not. The dialectic stopwords included words such as "اللي (Ally)" (meaning "that") and "مش (m$)" (meaning "not").

We used SOLR (ver. 4.0), which is built on top of Lucene, to perform all experimentation. We used the OKAPI-BM25 ranking

[3] http://members.unine.ch/jacques.savoy/clef/index.html

Table 3: Retrieval results for different runs

Run	*P@30*	*Significantly better than*
Norm	0.596	-
NormD	0.644	*Norm*
NormDS	0.624	-
Stem	0.673	*Norm*
StemD	0.694	*Norm, Stem*
StemDS	0.676	*Norm*
Char4g	0.619	-

formula with parameters K1=1.2 and b=0.76 [14]. We measured effectiveness using P@30, as in the TREC-2011 Microblog track, and used a paired 2-tailed t-test with p-value less than 0.05 to ascertain statistical significance.

5.2 Results and Discussion

Table 3 reports the results for all the runs on the full collection along with indication of which runs were statistically significantly better than others. The results suggest the following:

- Performing extended normalization led to statistically significant improvement whether we employed stemming or not.
- Unlike previously reported Arabic retrieval results (on a collection of 383k newswire stories): Using character 4-grams did not yield any improvement; and removing stopwords adversely affected retrieval effectiveness – though not significantly, which we suspect is due to the effect of removing stopwords on the length of documents
- Though stemming improved retrieval effectiveness, the difference between *NormD* and *StemD* was not statistically significant. Previously reported results in the literature showed that stemming led to statistically significant improvements. In our analysis we observed that:

- Stemming changed the intent of some queries. Consider the query "المتظاهرات في مصر – AlmtZAhrAt fy mSr" (the female protestors in Egypt). This query is related to incidents in which female protestors were attacked by security personnel, causing public uproar. When stemming was applied to the word for "the female protestors" (AlmtZAhrAt), the determiner and the feminine plural marker were removed leading to the singular masculine word for "protestor". Stemming led to 90 basis points drop in P@30. In other queries, removing the feminine plural or singular markers was beneficial, as in: "اشتباكات في مصر – A$tbAkAt fy mSr" (confrontations in Egypt), where the word for "confrontations" (A$tbAkAt) in Arabic is a feminine plural and is stemmed to the singular masculine form for "confrontation", and "الحركة المعارضة في البحرين – AlHrkp AlmEArDp fy AlbHryn" (opposition movement in Bahrain), where the word for opposition (AlmEArDp) is feminine singular and is stemmed to the singular masculine word for "an opposing person".
- Due to the large number of tweets and the limited number of retrieved tweets (only 30), most morphological forms are found often. Consider the Arabic words "والثورة (wAlvwrp)" (and the revolution), "الثورة (Alvwrp)" (the revolution), and "ثورة (vwrp)" (revolution), which appear in 15,593, 731,859, and 221,410 tweets respectively. In this typical example, though some morphological forms are more common than others, the least frequent form still appeared thousands of times.

Based on the results, we can glean that:

1. Performing improved letter normalization and elongated and shortened word handling is important.
2. Stemming leads to general improvements, though trained for MSA, but may change user intent. Investigation is required to know: when stemming is beneficial (or not); and how to combine both stemmed and unstemmed forms in ranking.
3. Stopword removal has unstable effect on retrieval effectiveness.

6. CONCLUSION AND FUTURE WORK

In this paper we focused on retrieval of Arabic microblogs. We presented issues that complicate retrieval and ways we can overcome some of these issues. We introduced new extended word normalization that includes improved letter normalization and elongated and shortened word handling. We also constructed a new stopword list that covers MSA as well as dialectic Arabic. Our experiments show that our extended normalization significantly improved upon state-of-the-art normalization, and using extended normalization can lead to results that are comparable to using stemming. Combining stemming with extended normalization leads to further improvement. We found that stopword removal often may lead to degradation in retrieval effectiveness, but the differences were not statistically significant.

For future work, stemming and stopword removal require further investigation for Arabic microblog retrieval tasks.

REFERENCES

1. K. Darwish, H. Hassan, O. Emam. (2005). Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval. *CASL workshop in ACL 2005*.
2. M. Diab. (2009). Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. *2nd Int. Conf. on Arabic Lang. Resources & Tools*.
3. P. Ferguson, N. O'Hare, J. Lanagan, O. Phelan, K. McCarthy. (2012). An Investigation of Term Weighting Approaches for Microblog Retrieval. *ECIR 2012*.
4. F. Gey, D. Oard. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. *TREC-2001*.
5. N. Habash. (2010). Introduction to Arabic Natural Language Processing (Synthesis Lectures on Human Language Technologies). *Morgan & Claypool Publishers*.
6. B. Han, T. Baldwin. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. *ACL-HLT 2011*.
7. L. Larkey, L. Ballesteros, M. Connell. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. *SIGIR 2002*.
8. D. Metzler, C. Cai. (2011). USC/ISI at TREC 2011: Microblog Track. *TREC-2011*.
9. N. Naveed, T. Gottron, J. Kunegis A. Alhadi. (2011). Searching microblogs: coping with sparsity and document quality. *CIKM-2011*.
10. D. Oard, F. Gey. (2002). The TREC 2002 Arabic/English CLIR Track. *TREC-2002*.
11. Z. Obukhovskaya, K. Pervyshev, A. Styskin, P. Serdyukov. (2011). Yandex at TREC 2011 Microblog Track. In TREC-2011.
12. I. Ounis, C. Macdonald, J. Lin, I. Soboroff. (2011). Overview of the TREC-2011 Microblog Track. *TREC-2011*.
13. O. Phelan, K. McCarthy, M. Bennett, and B. Smyth. (2011). Terms of a feather: content-based news recommendation and discovery using twitter. *ECIR 2011*.
14. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock- Beaulieu, M. Gatford. (1995). Okapi at TREC-3. *TREC-3*.
15. I. Subasic, B. Berendt. (2011). Peddling or Creating? Investigating the Role of Twitter in News Reporting. *ECIR-2011*
16. J. Teevan, D. Ramage, M. Morris. (2011). #Twittersearch: A comparison of microblog search and web search. *WSDM 2011*.
17. K. Versteegh. (1997). Dialects of Arabic. The Arabic Language. *Edinburgh University Press*.
18. Z. Wei, L. Zhou, B. Li, K.-F. Wong, W. Gao, K.-F. Wong. (2011). Exploring Tweets Normalization and Query Time Sensitivity for Twitter Search. *TREC-2011*.