Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 20–37

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

SAMAR: Subjectivity and sentiment analysis for Arabic social media[☆]

Muhammad Abdul-Mageed [a,b,*], Mona Diab [c], Sandra Kübler [a]

[a] *Department of Linguistics, Indiana University, 1021 E 3rd. St., Bloomington, IN 47405, USA*
[b] *School of Library and Information Science, 1320 East 10th Street, Bloomington, IN 47405, USA*
[c] *Department of Computer Science, School of Engineering & Applied Science, The George Washington University, Washington, DC, USA*

Received 16 August 2012; received in revised form 16 December 2012; accepted 13 March 2013
Available online 26 March 2013

Abstract

SAMAR is a system for subjectivity and sentiment analysis (SSA) for Arabic social media genres. Arabic is a morphologically rich language, which presents significant complexities for standard approaches to building SSA systems designed for the English language. Apart from the difficulties presented by the social media genres processing, the Arabic language inherently has a high number of variable word forms leading to data sparsity. In this context, we address the following 4 pertinent issues: how to best represent lexical information; whether standard features used for English are useful for Arabic; how to handle Arabic dialects; and, whether genre specific features have a measurable impact on performance. Our results show that using either lemma or lexeme information is helpful, as well as using the two part of speech tagsets (RTS and ERTS). However, the results show that we need individualized solutions for each genre and task, but that lemmatization and the ERTS POS tagset are present in a majority of the settings.

Keywords: Subjectivity and sentiment analysis; Morphologically rich language; Arabic; Social media data

1. Introduction

In natural language, *subjectivity* refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982) and, as such, it incorporates sentiment. The process of subjectivity classification refers to the task of classifying texts as either *objective* (e.g., *The new iPhone was released.*) or *subjective*. Subjective text can further be classified with *sentiment* or *polarity*. For sentiment classification, the task consists of identifying whether a subjective text is *positive* (e.g., *The Syrians continue to inspire the world with their courage!*), *negative* (e.g., *The bloodbaths in Syria are horrifying!*), *neutral* (e.g., *Obama may sign the bill.*), or, sometimes, *mixed* (e.g., *The iPad is cool, but way too expensive.*).

In this work, we address two main issues in subjectivity and sentiment analysis (SSA): First, SSA has mainly been conducted on a small number of genres such as newspaper text, customer reports, and blogs. This excludes, for example, social media genres, such as Wikipedia Talk Pages. Second, despite increased interest in the area of SSA,

[☆] This paper has been recommended for acceptance by R.K. Moore.

[*] Corresponding author at: Department of Linguistics, Indiana University, 1021 E 3rd. St., Bloomington, IN 47405, USA. Tel.: +1 812 855 6456. *E-mail address:* mabdulma@indiana.edu (M. Abdul-Mageed).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.03.001

only few attempts have been made to build SSA systems for *morphologically-rich languages*, i.e., languages in which a significant amount of information concerning syntactic units and relations is expressed at the word-level (Tsarfaty et al., 2010), such as Finnish or Arabic, cf. (Abbasi et al., 2008; Abdul-Mageed et al., 2011a; Mihalcea et al., 2007). Thus, we aim at partially bridging these two gaps in research by presenting an SSA system for Arabic social media genres as Arabic is one of the most morphologically complex languages (Diab et al., 2007; Habash et al., 2009). We present SAMAR, a sentence-level SSA system for Arabic social media texts. We explore the SSA task for four different genres: Synchronous chat, Twitter, Web discussion fora, and Wikipedia Talk Pages. These genres vary considerably in terms of their functions and the language variety employed. While the chat genre is mostly in dialectal Arabic, the other genres are mixed between Modern Standard Arabic (MSA) and dialectal Arabic to varying degrees.

1.1. Research questions

In the current work, we focus on investigating four main research questions:

- **RQ1:** How can morphological richness be treated in the context of Arabic SSA? To date most robust SSA systems have been developed for English, which has relatively little morphological variation. In such systems most of the features are highly lexicalized, hence a direct application of these methods would not be quite as successful for Arabic since a lemma in Arabic may be associated with hundreds if not thousands of variant surface forms. Accordingly, we need to investigate how to avoid data sparseness resulting from using lexical features without losing information that is important for SSA. More specifically, we characterize our problem in two spaces: the lexical space comparing simple lexeme tokenization with full lemmatization (lexemes vs. lemmas); abstracting away from the lexical form to the part of speech class, we investigate using two different POS tag sets for Arabic that encode a significant amount of morphological information.
- **RQ2:** Can standard features be effective for SSA when handling social media despite the inherently short texts typically used in these genres? In this prong of the research we investigate the impact of using two standard features frequently employed in SSA studies (Wiebe et al., 2004; Turney, 2002) on social media data that employes dialectal Arabic usage and the text inherently varying in length (i.e., the text being very short, e.g., in Twitter data). First, we investigate the utility of applying a UNIQUE feature (Wiebe et al., 2004) where low frequency words, below a certain threshold, are replaced with the token “UNIQUE”. Given that our data includes very short posts (e.g., twitter data has a limit of only 140 characters per tweet), it is questionable whether the UNIQUE feature will be useful or whether it replaces too many content words. Moreover, it should be noted that dialectal Arabic, to date, does not have a standardized orthography, therefore low frequency content words will be pervasive in social media genres since most of these genres employ dialectal Arabic. Second, we test whether a polarity lexicon that was extracted from a standard Modern Standard Arabic (MSA) newswire domain is useful for processing SSA for social media data.
- **RQ3:** How do we handle dialects in an SSA system for Arabic? For Arabic, there are significant differences between dialects on all levels of linguistic representation: morphology, lexical, phonology, syntax, semantics, and pragmatics. This difference is even more pronounced between the dialects and MSA. However, existing robust Arabic NLP tools such as tokenizers, Part of speech (POS) taggers, and syntactic parsers are exclusively trained on and for MSA newswire genres. Therefore we would like to measure the impact on SSA performance of explicitly modeling for dialectal usage.
- **RQ4:** Which features specific to social media can we leverage? We are interested in investigating the impact of using information that is typically present in social media (meta) data such as gender, author and document id information on SSA performance.

The remainder of the paper is organized as follows: In Section 2, we give an overview of the linguistic characteristics of Arabic that are important for our work; Section 3 describes the social media corpora and the polarity lexicon used in the experiments; In Section 4, we review related work; Section 5 describes the SSA system, SAMAR, used for the current research, as well as the features used in the experiments; Section 6 describes the experiments and discusses the results; In Section 7, we give an overview of the best settings for the different corpora, followed by a conclusion in Section 8.