Journal of King Saud University – Computer and Information Sciences (2014) **26**, 417–424

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Building an Arabic Sentiment Lexicon Using Semi-supervised Learning

Fawaz H.H. Mahyoub [a,b], **Muazzam A. Siddiqui** [a,*], **Mohamed Y. Dahab** [a]

[a] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
[b] *Faculty of Computer Sciences and Information Technology, Taiz University, Taiz, Yemen*

Available online 28 September 2014

KEYWORDS

Sentiment lexicon;
Sentiment analysis;
Arabic natural language processing;
Text mining;
Semi-supervised learning

Abstract Sentiment analysis is the process of determining a predefined sentiment from text written in a natural language with respect to the entity to which it is referring. A number of lexical resources are available to facilitate this task in English. One such resource is the SentiWordNet, which assigns sentiment scores to words found in the English WordNet. In this paper, we present an Arabic sentiment lexicon that assigns sentiment scores to the words found in the Arabic WordNet. Starting from a small seed list of positive and negative words, we used semi-supervised learning to propagate the scores in the Arabic WordNet by exploiting the synset relations. Our algorithm assigned a positive sentiment score to more than 800, a negative score to more than 600 and a neutral score to more than 6000 words in the Arabic WordNet. The lexicon was evaluated by incorporating it into a machine learning-based classifier. The experiments were conducted on several Arabic sentiment corpora, and we were able to achieve a 96% classification accuracy.

1. Introduction

Sentiment analysis is the process of determining a predefined sentiment from online texts written in a natural language with respect to a specific subject. The need for sentiment analysis is the product of a sudden increase in opinionated or sentimental texts in the form of blogs, reviews, and discussions (Pang and Lee, 2008). The idea of processing these comments or reviews has attracted many researchers in the field of text mining, with the aim of extracting a general opinion about one item or theme among the substantial amounts of unstructured data available on the Internet. In this paper, we present an Arabic sentiment lexicon that was developed by exploiting the semantic relations found in the Arabic WordNet. While there are several previous examples of using WordNet to build an English sentiment lexicon (Kim and Hovy, 2004; Esuli and Sebastiani, 2005, 2006), to the best of our knowledge, this is the very first attempt to build an Arabic sentiment lexicon using the Arabic WordNet. The Arabic WordNet is the Arabic version of WordNet and can be seen as a network with a collection of semantically similar words, called synsets, as nodes and a number of semantic and lexical relations as links between the synset nodes. We used a semi-supervised approach to propagate the sentiment scores from a small seed list of positive and negative

[*] Corresponding author.
E-mail addresses: fawazh7@gmail.com (F.H.H. Mahyoub), maasiddiqui@kau.edu.sa (M.A. Siddiqui), mdahab@kau.edu.sa (M.Y. Dahab).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.003

words in the Arabic WordNet. We devised an algorithm that identified the nodes in the Arabic WordNet that contain the words in the seed list and iteratively spread the scores of these words to the neighboring nodes until the entire network was reached. The score for each term was represented as a triplet containing a positive, negative and neutral score. Each of these constituent scores in the triplet for a term was represented as positive numerical values. The scheme is somewhat similar to how scores are represented in the SentiWordNet, but in our case, the scores were unnormalized, i.e., the positive, negative and neutral scores of the term do not sum to one.

The main contribution of this work is the development of an Arabic sentiment lexicon containing 7.5 K terms by exploiting the relations available in the Arabic WordNet. In addition to the sentiment scores, the lexicon also contains the part of speech tag of each term and its diacritized form for lexical disambiguation. For some of the terms, the gloss containing the term definition is also available.

The remainder of this paper is structured as follows: The next section briefly describes the Arabic WordNet. In Section 3, we present the previous major approaches to developing a sentiment lexicon. In Section 4, we describe the development of an Arabic sentiment lexicon. In Section 5, we evaluate the proposed algorithm. Finally, Section 6 is devoted to conclusions and future work.

2. What is the Arabic WordNet?

WordNet is a lexical database of the English language. Unlike a dictionary, the words, including nouns, verbs, adjectives and adverbs, are grouped into sets of synonyms called synsets. These synsets are related to each other through different semantic and lexical relations; hence, the WordNet can be viewed as a directed graph (Fellbaum, 1998). The Arabic WordNet is the Arabic version of the English WordNet. The Arabic WordNet database structure is composed of four principal entity types: item, word, form and link. Items are conceptual entities, including synsets, ontology classes and instances. A word entity is a word sense. A form is a special form that is considered as dictionary information. Links are relations between synsets. They are classified according to the part of speech (POS) of the related synsets (verb, noun, adjective, and adverb) or according to their type (lexical, semantic and lexico-semantic relations). Table 1 presents WordNet and Arabic WordNet statistics (WordNet 3.0 database statistics), (Fellbaum et al., 2006). Table 2 shows different relations in the Arabic WordNet according to their classification type (Mahdi Boudabous et al., 2013).

Table 1 WordNet and Arabic WordNet database statistics.

POS	AWN		PWN	
	Word forms	Synsets	Word forms	Synsets
Noun	15,890	7,960	117,798	82,115
Verb	6,084	2,538	11,529	13,767
Adjective	1,243	661	21,479	18,156
Adverb	264	110	4,481	3,621
Total	23,481	11,269	155,287	117,659

3. Related work

Although plenty of research is available on building sentiment lexicons in English and other languages, Arabic has yet to receive the attention it deserves by researchers in this field. In this section, we will present the most notable studies on building English sentiment lexicons and previous attempts to build Arabic sentiment lexicons. In addition, we will also cover studies that claim language independence.

Hatzivassiloglou and McKeown (1997) developed an algorithm for predicting the orientation of an adjective. Turney and Littman (2002) proposed a method to determine a document's polarity. The method involves issuing queries to a Web search engine. The approach targets adjectives and adverbs; therefore, it relies on the existence of a huge POS-tagged corpus, which is a rarity for the Arabic language. The available POS taggers are not fully qualified to identify all parts of speech and are not able to distinguish between different sentence types (Farra et al., 2010). Lexical resources, such as WordNet (Fellbaum, 1998), are used in Kim and Hovy (2004), Esuli and Sebastiani (2005, 2006), Kamps et al. (2004). These studies started with

Table 2 Arabic WordNet relation classification (Mahdi Boudabous, 2013).

Type	Relation	Example	Frequency
Semantic relations	Has hyponym	ماء, حَليب has hyponym شَرَاب	9352
	Has holo part	فرنسا, إسبانيا has holo part أوروبا	697
	Has subvent	بَلَعَ has subvent أَكَلَ	128
	Has instance	مرسيليا has instance مدينة	1067
	See also	عَرَفَ See also إكْتَشَفَ	192
	Causes	تذَكَّرَ Causes ذَكَّرَ	75
	Has holo member	الاتحاد الأوروبي has holo member فرنسا	334
	Verb group	سَمَحَ – أَجَازَ	152
	Region term	حروب اليونان Region term البلقان	35
	Category term	عقيد, لواء catégorie termes جيش	548
	Has holo made of	ماء has holo made of هيدروجين, أكسجين	60
	Be in state	قادر Be in state إمْكَان	83
	Usage term	أسبرين Usage term اسم تجاري	3
Lexical relations	Near synonym	انسان near synonym فرد	122
	Near antonym	أم Near antonym أب	722
Lexico-semantic relations	Related to	إحْتِفَاظ related to إحْتَفَظَ	4774
	Has derived	والدَان has derived أَبَوِيّ	178