Int J Speech Technol (2016) 19:857–867
DOI 10.1007/s10772-016-9376-y

Word sense disambiguation for Arabic text using Wikipedia and Vector Space Model

Marwah Alian[1] · Arafat Awajan[2] ·Akram Al-Kouz[2]

Received: 14 July 2016 / Accepted: 26 September 2016 / Published online: 4 October 2016 © Springer Science+Business Media New York 2016

Abstract In this research we introduce a new approac h for Arabic word sense disambiguation by utilizing Wikipedia as a lexical resource for disambiguation. The nearest sense for an ambiguous word is selected using Vector Space Model as a representation and cosine similarity between the word context and the retrieved senses from Wikipedia as a measure. Three experiments have been conducted to evaluate the proposed approach, two experiments use the first retrieved sentence for each sense from Wikipedia but they use different Vector Space Model representations while the third experiment uses the first paragraph for the retrieved sense from Wikipedia. The experiments show that using first paragraph is better than the first sentence and the use of TF -IDF is better than using abstract frequency in VSM. Also, the pr oposed approach is tested on English words and it gives better results using the first sentence retrieved from Wikipedia for each sense.

Keywords Arabic word sense disambiguation · Disambiguation resource · Vector space model · Wikipedia

Marwah Alian
Marwah2001@yahoo.com

Arafat Awajan
awajan@psut.edu.jo

Akram Al-Kouz
akram@psut.edu.jo

[1] Hashemite University, Zarqa, Jordan

[2] Princess Sumaya University for Technology, Amman, Jordan

1 Introduction

One of the most difficult problems in Natural Language Processing (NLP) is the capability to identify what a word means with respect to its context. The technique that is used to find the appropriate sense of a word with an ambiguous meaning considering its context is called Word Sense Disambiguation (WS D). (Ide and Ve´ronis 1998; Navigli 2009). It is ubiquitous across all languages but it has greater challenges in Semitic languages like Arabic. WSD is considered as an AI -complete problem (Mallery 1988) and it is required in several applications such as machine translation (Carpaut and Wu2005; Chan et al. 2007), Information retrieval (Schu¨tze and Pedersen 1995; Stokoe et al. 2003) and information extraction (Jacquemin et al. 2002).

A word may denote different meanings in two different sentences because of ambiguity features in human languages. For example, in English the word bass may have a different meaning according to the context in which it comes into view (Navigli 2009) such as:

I can hear bass sounds.

They like grilled bass.

The word bass in the first sentence mean s musical instruments while in the second sentence means a fish. Furthermore, world knowledge is required in order to identify that the sense of ‘bass’ is a fish and not a musical instrument in the second sentence. This is because one could grill a fish not an instrument.

In order to measure the similarity between two texts, which are represented mathematically using Vector Space Model (VSM), if the text that contains the sense of an ambiguous word and the query text have a similar column