Arabic Word Sense Disambiguation - Survey

Marwah Alian
Hashemite University
marwah2001@yahoo.com

Arafat Awajan
Princess Sumaya University for Technology
awajan@psut.edu.jo

Akram Al-Kouz
Princess Sumaya University for Technology
akram@psut.edu.jo

Abstract— **One of the central challenging and most difficult problems in Natural Language Processing is the capability to identify what a word means with respect to a context in which it comes into view. This problem is called Word Sense Disambiguation (WSD). It is ubiquitous across all languages but it has greater challenges in Semitic languages like Arabic language. In this paper we present what researches have been done to solve the problem of Arabic word sense disambiguation.**

Keywords— Word Sense Disambiguation; Natural Language processing; Arabic Word Sense Disambiguation.

I. INTRODUCTION

The mechanism that is used to find the appropriate sense of a word that has ambiguous meaning considering its context is called Word Sense Disambiguation (WSD) [*1*] [*2*]. This technique is used in several applications, for example in machine translation [*3*] [*4*], information extraction [*5*] and Information retrieval [*6*] [*7*] .

Some words are called ambiguous words where these words have different meaning according to their context. For example:

Notre Dame stands in the very heart of Paris.

His heart moved him to help the needy.

In the first sentence the word heart means center while in the second sentence it means source of feelings.

In the survey of [*8*] a classification is used for the methods of word sense disambiguation in English language. These classes are: Knowledge-based, Supervised and Unsupervised methods where each class has a number of methods that are use.

Knowledge-based approaches use dictionaries while the supervised approaches are based on hand labeled data that are typically lexical samples. The unsupervised approaches use the information found in un-annotated corpora to distinguish the word meaning [*8*].

Several researches have been introduced to solve the ambiguity of words in many languages, but it is limited in Arabic and there is no survey for Arabic word sense disambiguation AWSD. Therefore, the aim of this research is to represent what have been introduced to solve AWSD using similar classes to that introduced in [*8*] with the addition of a Hybrid class. However, supervised approaches are rarely used for Arabic word sense disambiguation because of the lack of Arabic standard annotated corpora. So, this class is not included in the survey.

This paper is organized as follows: section II gives a brief history of what have been proposed during last years in WSD for English language. In section III, a review for methods used for AWSD is presented while discussion and future work are in section IV.

II. WORD SENSE DISAMBIGUATION IN ENGLISH

In [*9*] a genetic algorithm (GA) for Word Sense Disambiguation is introduced as well as a weighted genetic algorithm. In order to identify words senses, WordNet is used then GA is performed to maximize similarities with respect to sematic in the set that is obtained from WordNet. In weighted GA for Word Sense Disambiguation (WGWSD) they use averaging crossover and random mutation. The proposed algorithms were experimented on SemCor files and the comparison of their results with other works presents a better disambiguation precision for WGWSD but it degrades when it is compared with the results of the work presented by [*10*]. This research did not consider disambiguation of verbs and adjectives, it considers just nouns.

In [*11*] the researchers focus on the modal verb level by studying the word sense disambiguation of the verb "may". They construct a neural network with back propagation model depending on the sense analysis, class of modality and the context functionality of the verb "may". The study represents a 78% disambiguation accuracy since the system has a high rate of fault tolerant capability as well as self-adaptivity.

In the work of [*12*], the researchers proposed a new approach for WSD using machine translation from Arabic language to English language. It depends on the richness of Arabic language in morphology and translation with WSD involved. The proposed approach uses Naïve Classifier with some modification because Arabic language has features have to be taken into consideration as well as the large size of the corpus. They use precision and applicability to measure the performance and the experiment shows that the proposed approach gives better results for long query terms but it degrades in short query. Two classifiers are constructed and the results for the new approach give 68% precision for Query term with topic context classifier and 93% precision for query term with feature inflectional form classifier.

In the study of [*13*] a comparison between methods that is mainly used in WSD is presented then a proposed method

978-1-5386-0527-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICTCS.2017.23