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

is introduced which depends on the computation of similarity for multi-level sentence in Vector Space Model. The new method is expected to increase the accuracy to some extent and has the advantage of removing the problem of VSM that is bag of words.

While the research of [*14*] introduce an enhancement on the graph based algorithm where In-Degree and the similarity measure that compute verb to verb similarity is modified. This is an unsupervised technique in which all tokens are connected to meanings that are relevant to their context from a lexical resource. In the experiment, data sets from three standards were used on three versions of WordNet. These standard data sets are SENSEVAL2, SENSEVAL3 and SEMEVAL. The results of this work were the best monolingual unsupervised results on the standard data sets.

In [*15*] they represent an unsupervised approach to choose the best sense of an ambiguous word. They introduce a definition for a probabilistic Latent Semantic analysis depending on WSD system. In this system tagged senses are not needed to train the system because the system does not depend on the language. This approach produce a percentage of 83% for accurate selection in English and 74% accuracy for Hindi language and the performance of this approach is increased with the use of WordNet.

The study of [*16*] proposes an unsupervised algorithm for Word Sense Disambiguation based on two versions of Lesk algorithm. In this new algorithm the computation of similarity and overlapping between the context in which the word exists and the definition of its senses in a semantic space that is distributed. For sense inventory, BabelNet is selected and for evaluation, SemEval-2013 is chosen. The results show that the proposed algorithm gives better results than the two versions of Lesk algorithm.

While in the study of [*17*], an update on Yarowsky algorithm is suggested as a solution for Word Sense Disambiguation in Quranic translation. In this research, they use a data set of Quranic content and the three IR evaluation metrics to evaluate the retrieving efficiency for the updated method. The process begins by preprocessing then two lists are constructed for every word that gives two senses each list contains examples from the dataset for the word with the list's sense. By experiment, they got 77% of f-measure and they consider it as competitive result in Word Sense Disambiguation, but it is a weak result if it is compared to the results of Yarowsky because of less extracted examples from Quran.

III. ARABIC WORD SENSE DISAMBIGUATION

A. KNOWLEDE BASED AWSD

Most of the researches introduced to solve AWSD use Knowledge-based approach [*18*] [*19*] [*20*] [*21*] [*22*] [*23*].

For example, the study of [*18*] introduce a system that is independent on the language and based on a self-expansion mechanism. It is based on replacing every term in the original data set with a set of relevant terms according to some calculations then clustering is performed on the expanded version. In this research Kstar clustering [*24*] was utilized in order to produce all possible senses for the ambiguous words and a partial tokenization is used because they kept the Arabic "Al" joint to the words. The system is experimented on both Arabic and English languages.

Also in [*19*], an evaluation for the variants of the Lesk algorithm was conducted for AWSD. They used the dictionary and perform the original Lesk algorithm [*25*] Also, they introduced and experimented modifications on Lesk algorithm. The experiments were performed using free tools while the data set were built based on several resources. They used similarity measures to identify how two concepts in Arabic Wordnet are similar. In Arabic work, the rate of precision was the best at window size equals to three words for original Lesk algorithm while it was at widow size of two words for modified Lesk algorithm.

While in the work of [*20*], a new approach was proposed to solve AWSD problem using Genetic Algorithm (GA), named as GAWSD. They tested their approach using a sample text in Arabic then they compared with naïve Bayes classifier. The proposed approach gave better results than naïve classifier. In this work the sample was small so it needs to be extended to a large Arabic corpus to check the performance.

In [*21*] the researchers propose a new method for AWSD they utilize both the global and local context of an ambiguous word where the correct sense was considered as the one with a closer semantic similarity to both local and global context where local context is specified by the neighborhood of an ambiguous word, and the global context is specified by the whole text. They claim that their proposed approach provided an accuracy of 74%.

A comparative study in [*26*] aims to present the possibility of using the Rocchio classifier to solve the Word Sense Disambiguation problem. The Rocchio classifier is tested and a comparison between its performance and the performance of three supervised approaches for word sense disambiguation; those approaches are: the Most Frequent Sense (MFS), Naïve Bayesian Classifier (NBC) and the Support Vector Machine (SVM). Rocchio classifier shows promising results as a supervised approach. It shows better results than other classification with 88% overall accuracy and it decreases the error by more than 14% when it is compared to NBC.

In the study of [*27*], three supervised algorithms for Arabic Word Sense Disambiguation are experimented and compared. These tested algorithms are; the Naïve Bayes algorithm, the decision list and the k nearest neighbor. A sample of fifty Arabic ambiguous words is used in the experiment, where the k nearest neighbor outperform the two algorithms. It is concluded from this experimental study that the supervised algorithms require tagged data with an important amount in order to get satisfying results.

In the work of [*28*] they introduce a method for disambiguating Arabic words using Wikipedia as a lexical resource. They implemented the sentences by Vector Space Model then compute the similarity using cosine similarity and the closest context for an ambiguous word is selected.

They test the use of the frequency of words and the TF-IDF weight for the words in the VSM when computing the similarity and the results show that the use of TF-IDF VSM is better than using raw frequency.

However, they extend their work in [*23*] and test their approach for English words and compare the results with disambiguating Arabic words using Wikipedia their experiments show that the best results are given when retrieving the first paragraph from Wikipedia texts for each sense.

B. UNSUPERVISED AWSD

The first system that handles word sense disambiguation for Arabic words was an unsupervised method that was introduced by [*29*] where an unsupervised mechanism is used based on the observation that words which have similar translation usually have similar dimension of meaning.

Their method adopt that a correct sense is strengthening by the semantic similarity of words sharing identical dimension of meaning. They utilize English WordNet in order to get words senses then English-Arabic corpus is used for translation. Their results were comparable to other unsupervised systems.

In [*30*] the researchers test the expansion of a query using interactive Word Sense Disambiguation (WSD) in search engine and they compare it with the query without expansion. In their study they concerned only with automatic information retrieval systems. They expand a query terms by the addition of more specific synonyms and test the searching results. The results of their experiment demonstrates that the expansion of a query terms will narrow the search and make it very close to the targeted request and it increases precision and recall. While the expansion of a query using more general synonyms decrease the precision.

While in [*31*], the researchers present a new technique for the disambiguation of Arabic language by using weighted directed graph. It is a semi-supervised method where the disambiguation procedure depends on matching the semantic tree with the original sentence tree and to detect the closest semantic tree, a score measure is used. The performance of the proposed technique is tested and compared with other works in AWSD such as supervised, unsupervised and knowledge base methods. The number of ambiguous words that were included in the experiment is fifty and the performance was measured under the number of nodes in the semantic tree. The results of this research demonstrate that the performance increases for semantic trees that have 500 nodes while F-score goes upward and being stable for semantic trees with 2,000 to 3,000 nodes. Their technique reaches a high recall and precision that is 83%.

Also in the work of [*32*], a fuzzy logic based technique is proposed to build a new classifier in order to be utilized in Arabic Word Sense Disambiguation. In this work, two fuzzy logic classifiers are constructed and compared with a Naïve Byes classifier. The classifiers are used to identify the most possible senses for a word that has ambiguity. They identify a list of ambiguous words consist of ten ambiguous words; they collect them from other researches handling the same problem. They argue that fuzzy logic considers the overlapping through different senses and that fuzzy logic deals with ambiguity and vagueness. Their experimental results show that fuzzy classifier gives more accurate results.

However, In the study of [*22*], the researchers utilize both English WordNet and Arabic WordNet depending on machine translation for terms and they select the closest concept for an ambiguous word using the relationships between the ambiguous word and the different concepts in local context. In their experiments they use different machine learning and feature selection techniques to evaluate their method. The results of their system show that the proposed approach outperforms other techniques for Arabic word sense disambiguation.

C. HYBRID METHODS FOR AWSD

In the study of [*33*], a new system is developed using hybrid technique that combine Lesk algorithm with other information retrieval methods. They use Croft, Latent semantic analysis, Harman, and Okapi methods with Lesk algorithm to determine which sense is the highly related sense to an ambiguous word. The estimated value of closeness between the context in which an ambiguous word appeared and the several contexts of every meaning of the word. They construct a database for different contexts for every sense of ambiguous word, meaning and synonyms in order to be utilized by the system. The system gives 73% correct determination of ambiguous words in a sample of ten ambiguous words.

Also, in [*34*] a hybrid technique is proposed for Arabic Word Sense Disambiguation. The hybrid technique integrated unsupervised and knowledge-based methods in order to give a correct word meaning. In this work, texts that have ambitious words are pre-processed and extraction is done for pertinent words then an algorithm is applied for context matching which gives a score for how the context of use is closest semantically to the original sentence. Matching algorithm associates a score for the relevant senses of the ambiguous word. This hybrid technique is distinguished by finding signatures, rooting and applying string matching algorithm. The results of this technique were compared with Naïve Bayes Classifier for AWSD and with the unsupervised approach. The hybrid technique outperforms other techniques in term of precision.

IV. DISCUSSION AND FUTURE WORK

In this research a review for the researches in the field of word disambiguation is presented for both English and Arabic languages.

The researches in Arabic language use different data sets with limited size and they are not available which make it difficult to consider their results. And the use of Arabic WordNet has several problems such as noise, precision and limited coverage compared to English WordNet.

Knowledge based approaches provide higher precision but they their performance depend on dictionary definitions while unsupervised approaches do not need sense annotated corpora but their performance is less than knowledge based

approaches. However the Hybrid approaches try to benefit from their component methods and give better results.

As a future work Arabic Wikipedia can be used as a resource for disambiguation to enrich the Arabic WordNet then the similarity can be measured between the retrieved texts from Wikipedia and the tested text which contains ambiguous word in order to assign the appropriate sense for an ambiguous word.

REFERENCES

[1] Ide, N., Véronis, J., "Word Sense Disambiguation: The State of the Art," *Computational Linguistics.*, vol. 24, no. 1, pp. 1-40, 1998.

[2] Roberto Navigli, "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1-69, 2009.

[3] Carpaut, M., and Wu, D., "Word Sense Disambiguation vs. Statistical Machine Translation," in *the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005, pp. 387-394.

[4] Chan, Y., Ng, H., and Chiang, D., "Word Sense Disambiguation Improves Statistical Machine Translation," in *45rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007, pp. 33-40.

[5] Jacquemin, B., Brun, C., and Boux, C., "Enriching a Text by Semantic Disambiguation for Information Extraction," in *the Workshop on Using Semantics for Information Retrieval and Filtering in the 3rd International Conference in Language Resources and Evaluation (LREC)*, 2002.

[6] Schütze, H., and Pedersen, J. , "Information Retrieval Based on Word Senses," in *Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 1995, pp. 161-175.

[7] Stokoe, C., Oakes, M., and Tait, J., "Word Sense Disambiguation in Information Retrieval Revisited," in *the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 159-166.

[8] Alok Ranjan Pal, Diganta Saha, "Word Sense Disambiguation: A Survey," *International Journal of Control Theory and Computer Modeling (IJCTCM)*, vol. 5, no. 3, 2015.

[9] ChunHui Zhang, Yiming Zhou,Trevor Martin, "Genetic Word Sense Disambiguation Algorithm," in *Second International Symposium on Intelligent Information Technology Application, IEEE*, 2008, pp. 123-127.

[10] P. Rosso, F. Masulli, D. Buscaldi. , "Word Sense Disambiguation combining Conceptual Distance, Frequency and Gloss.," in *International Conference on Natural Language Processing and Knowledge Engineering*, 2003, pp. 120-125.

[11] Jianping YU, Jian ZHANG, "Word Sense Disambiguation of the English Modal Verb May by Back Propagation Neural Network," in *International Conference on Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08.*, 2008, pp. 1-6.

[12] F. Ahmed ,A. Nurnberger, "Arabic/english word translation disambiguation approach based on naive bayesian classifier," in *International Multiconference on Computer Science and Information Technology IMCSIT*, 2008, pp. 331-338.

[13] "[16] Zhang Zheng, Zhu Shu, 2009, A New Approach to Word Sense Disambiguation in MT System," in *World Congress on Computer Science and Information Engineering*, 2009, pp. 407-411.

[14] Weiwei Guo, Mona T. Diab, "Improvements to Monolingual English Word Sense Disambiguation," in *NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 2009, pp. 64–69.

[15] Gaurav S Tomar, Manmeet Singh, Shishir Rai, Atul Kumar, Ratna Sanyal, Sudip Sanyal, "Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation," *IJCSI International Journal of Computer Science Issues*, vol. 10, no. 5 , pp. 1694-0784, 2013.

[16] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, "An Enhanced LeskWord Sense Disambiguation Algorithm through a Distributional Semantic Model," in *COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1591–1600.

[17] Omar Jamal Mohamed, Sabrina Tiun, "Word Sense Disambiguation Based on Yarowsky Approach In English Quranic Information Retrieval System," *Journal of Theoretical and Applied Information Technology*, vol. 82, no. 1, pp. 163-171, 2015.

[18] David Pinto, Paolo Rosso, Yassine Benajiba, Anas Ahachad, Héctor Jiménez-Salazar , "Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach," in *7th Conference on Language Engineering of the Egyptian Society of Language Engineering-ESOLE*, 2007, pp. 235-245.

[19] A. Zouaghi, L. Merhbene, M. Zrigui, "Word Sense disambiguation for Arabic language using the variants of the Lesk algorithm," in *WORLDCOMP'11*, 2011, pp. 561-567.

[20] Mohamed El Bachir Menai, Wojdan Alsaeedan, "Genetic algorithm for Arabic word sense disambiguation, ," in *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE*, 2012, pp. 195-200.

[21] Nadia Bouhriz, Faouzia Benabbou, El Habib Ben Lahmar, "Word Sense Disambiguation Approach for

Arabic Text," *International Journal of Advanced Computer Science and Applications (IJACSA)* , vol. 7, no. 4, pp. 381-385, 2016.

[22] Meryeme Hadni, Said El Alaoui, Abdelmonaime Lachkar, "Word Sense Disambiguation for Arabic Text Categorization," *The International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 215-222, 2016.

[23] Marwah Alian,Arafat Awajan, Akram Al-Kouz, "Word sense disambiguation for Arabic text using Wikipedia and Vector Space Model," *International Journal of Speech and Technology* , vol. 19, pp. 857–867, 2016.

[24] John G. Cleary, Leonard E., "Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure," in *12th International Conference on Machine Learning*, 1995, pp. 108-114.

[25] M. Lesk. 1986, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone," in *SIGDOC '86*, 1986.

[26] Soha M. Eid, Almoataz B. Al-Said, Nayer M. Wanas, Mohsen A. Rashwan, Nadia H. Hegazy, "A Comparative Study of Rocchio Classifier Applied to supervised WSD Using Arabic Lexical Samples," in *10th Conference on Language Engineering; CLE'2010, the Egyptian Society of Language Engineering (ESOLE)*, Cairo, 2010.

[27] Laroussi Merhben, Anis Zouaghi, and Mounir Zrigui, "Lexical Disambiguation of Arabic Language: An Experimental Study," *The Journal Polibits* , vol. 46, pp. 49- 54, 2012.

[28] Marwah Alian, Arafat Awajan, Akram Al-Kouz, "Arabic Word Sense Disambiguation Using Wikipedia," *International Journal of Computing & Information Sciences*, vol. 12, no. 1, September 2016.

[29] M. Diab, P. Resnik, "An unsupervised method for word sense tagging using parallel corpora," in *the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, 2002, pp. 255–262.

[30] R. Al-Shalabi, G. Kanaan, M. Yaseen, B. Al-Sarayreh, N. A. Al-Naji,2009, "Arabic query expansion using interactive word sense disambiguation," in *2nd Int. Conf. on Arabic Language Ressources and Tools, The MEDAR Consortium*, Cairo, Egypt, 2009.

[31] Laroussi Merhbene, Anis Zouaghi, Mounir Zrigui, "A Semi-Supervised Method for Arabic Word Sense Disambiguation Using a Weighted Directed Graph," in *International Joint Conference on Natural Language Processing*, 2013, pp. 1027-1031.

[32] Madeeh Nayer El-Gedawy, "Using Fuzzifiers to Solve Word Sense Ambiguation in Arabic Language," *International Journal of Computer Applications* , vol. 79, no. 2, pp. 0975 – 8887, 2013.

[33] Laroussi Merhbene,Anis Zouaghi, Mounir Zrigui, "Ambiguous Arabic Words Disambiguation: The results ," in *Student Research Workshop, RANLP*, Bulgaria, 2009 , pp. 45–52.

[34] Anis Zouaghi, "A Hybrid Approach for Arabic Word Sense Disambiguation," *International Journal of Computer Processing Of Languages*, vol. 24, no. 2, pp. 133–151, 2012.