Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Word Sense Disambiguation (WSD) and Information Retrieval (IR): Literature Review

Boshra F. Zopon Al_Bayaty[*]
University of Mustansiriyah – Iraq
Bharati Vidyapeeth Deemed University, India

Dr. Shashank Joshi[*]
College of Engineering
Bharati Vidyapeeth Deemed University, India

Abstract— Since the filed of word sense disambiguation and information retrieval has long history, the literature on this field is growing fast through the recent years. In this literature we tried collect some of research and studies about the relationship between the word sense disambiguation and information retrieve. Cover all the studies in this field is so difficult, for those we reading different papers and submitted as literature review and we hope that we can add little of contribution in this field.

Keywords: word sense disambiguation (WSD), Information retrieves (IR).

I. Introduction

In recent years WSD become one of central challenges in (NLP), and there are a lot of tasks in this field require and need disambiguation. Words almost have many meaning (multiple), and WSD is the task to determine the proper meaning of word and use it in particular context. WSD is considered as classification task because the word senses can be the classes and the automatic classification approach is used to identify and assign each occurrence of the word to classes from external knowledge sources. The information retrieval (IR) (Manning et al.,2008)is considered as a successful applied domain of (NLP), Its always deals with access to information such as Web pages, documents, pictures, video, and so on. There are many attempts to enhance the speed of information retrieval from the internet by use automated word sense disambiguation.

II. LITERATURE REVIEW

1. **Antonio Jimeno-Yepes, Alan R. Aronson, National Library of Medicine, 8600 Rockville Pike, Bethesda, 20894, MD, USA, "Self-training and co-training in biomedical word sense disambiguation", 2011.**

   These authors described the attempting to determine the sense of ambiguous words (WSD) as intermediate stage. They submitted the results about two algorithms of semi supervised learning with baseline and applied them on biomedical words. They achieved better results, and they put plan for their algorithms to apply it on the set of ambiguous words called MSH WSD set.

2. **Anna Bryniarska, "The Paradox of the Fuzzy Disambiguation in the Information Retrieval", Electrical and Computer Engineering, Opole University of Technology, Opole, Poland. 2013.**

   In this paper the author has talked about the existing methods in data mining, which involve different domains like word sense disambiguation and information retrieval, fuzzy sets theory and so on. He submitted new approach called paradox of the Fuzzy disambiguation, which lied on the fact that can gain voluble knowledge from due to fuzzy data and the expert's knowledge. The author also reported that his research will be useful to extend the software or to create new kind of search engine.

3. **Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, and Aijun An, "Cross-Lingual Word Sense Disambiguation for Languages with Scarce Resources" Department of Computer Science and Engineering, York University, Canada, 2013.**

   In this paper the authors have talked about advantages of English sense disambiguaters available in both in Wikipedia and their new a cross lingual proposed a approach in Persian text using FarsNet, and they achieved in their experiment results about a 28% enhancement in a curacy and approved the cross lingual method achieved best performance than the knowledge based method which is applied to Persian texts directly. The paper was for tested the Novel idea for Cross – lingual WSD in feasibility and other items. The authors suggested to achieve good performance in accuracy the will use sense relate with another English sense tagger as a step for future work.

4. **Caden Howell, Information Retrieval, CSC 575, Dr. Joe Phillips, "Information Retrieval Database with WordNet Word Sense Disambiguation", 2009.**

   In this paper, the author constructed information retrieval system and he used WordNet as a powerful tool in enhancement. And he have took index at the base of the system is simulated a MySQL database. Basically the database involved four tables: Document, Term, TermDocument, and TrialIndex. The authors recorded their notes about the relationship between different tables in the database of this system. For future work, the author suggested use standard known corpus in order achieve relationship between the WordNet indices with some objectivity.

5. Christopher M. Stokoe, Prof. John Tait, "Automated Word Sense Disambiguation for Internet Information Retrieval", 2002.

In this paper the authors have provided detailed discussion about their project and they talked about the use of automated disambiguation to achieve enhancement in information retrieval performance from the internet. They achieved a small "0.0003%" increase in R-accuracy. The authors proved the increase in training data can be providing good performance. They achieved about 39.9% error rate of their disambiguation methodology and did not have any an errors in information retrieval.

6. Christopher Stokoe, Michael P. Oakes, John Tait, The University of Sunderland Informatics Centre, "Word Sense Disambiguation in Information Retrieval Rvisited", 2003.

In this paper, the authors have explained a system for assign sense based information retrieval. In their disambiguation system strategy used a combination of high precision techniques and sense frequency statistics in order to get minimum of erroneous disambiguation on retrieval performance. The compared their experimentation with Schutze and Pederson, they achieved good results through made a word to be tagged with up to three possible word senses and combining word and sense ranking. They noted that with an accuracy 62% their experimentation achieved absolute progress of 1.37% and a relative increase over TF*IDE of 45.9%. The first reason of their experimentation was reduce the impact of the erroneous disambiguation. The author in their paper conclusion presented some key ideas like little bit disambiguation can provide increased accuracy in information retrieval also using WSD in IR very useful only within special kinds.

7. Francis de la C. Fernández REYES, Exiquio C. Pérez LEYVA, Rogelio Lau FERNÁNDEZ Instituto Superior Politécnico, José Antonio Echeverría, Marianao, Cuba, "Word Sense Disambiguation in Information Retrieval", 2009.

The authors in this paper have been proposed algorithm by combine positive features of the unsupervised approach by used of a classifiers set. In their method the authors defines a target word as a set of sense vectors, which include all kind of linguistic information in each vector. The authors continued in their work and shown that the supervised approaches always achieved best result than unsupervised one.

8. Hwee Tou Ng, "Does Word Sense Disambiguation Improve Information Retrieval?" Department of Computer Science National University of Singapore, 2012.

In this paper, the author have question (Does Word Sense Disambiguation Improve Information Retrieval?). The author has talked about semantic annotation of word senses improve information retrieval and others doesn't. He presented some of earlier work related with some researchers like Krovetz and Croft whom employed manual sense annotation and was reported bad results and shown the information retrieval doesn't improve performance. Also he talked about some researches works whom reported good results like Schutze and pedersen in their a well known study that "Unsupervised WSD improve IR performance". And Gonzalo, Mihalcea works whom achieved good results, and so on. This paper gave some thoughts on this question about the relationship between WSD and information retrieval.

9. Mark Sanderson, "word sense disambiguation and information retrieval", department of computing science, university of Glasgow, United Kingdom, 1995.

The author in this paper has talked about the belief of the word sense disambiguate is a cause of poor performance in information retrieval system. The done retrieval experiment and he reported that the disambiguator should be able to resolve word senses with high accuracy manner. The author has concluded that with an accuracy of 75%, the retrieval performance worse than performance using the ambiguous collection, and he achieved similar performance to the ambiguous collection with disambiguation at 90% accuracy. For this reasons the author suggested that before the practical use of general tools for computational linguistics tasks its need to operate at least 90% accuracy.

10. NITIN INDURKHYA FRED J. DAMERAU, "HANDBOOK OF NATURAL LANGUAGE PROCESSING SECOND EDITION", by Taylor and Francis Group, LLC, 2010.

In this handbook edition, the authors has talked about the application of WSD in information retrieval (IR), and shown it may be more benefit to map the word into separate clearly terms. This handbook included many experiments and can consider as one of a very good book in natural language processing field. The book involved many chapters deal with experiments of NLP field and the section 19.3 explain different IR approaches. Also section 19.4 showed in detailed the evaluation methodology that deal with information retrieval and doing compare between search methodologies.

11. Pushpak Bhattacharyya, Mitesh Khapra, Indian Institute of Technology Bombay, India, Word Sense Disambiguation, Chapter 2, 2010.

In this chapter of thesis, the authors have described the basic concepts of word sense disambiguation (WSD) and the methods of this field. And they focused on the different current methods for word sense disambiguation, and the have done comparison of different methods in this field. The authors submitted a greedy neural network inspired algorithm for word sense disambiguation and they made comparison with other algorithms performance, and they found that their experiments suggest for domain- specific WSD, the choosing the most frequent sense of a word its work like any state – of – the art algorithm.

12. **Ping Chen and Chris Bowes, University of Houston-Downtown, Wei Ding and Max Choly, University of Massachusetts, Boston, "Word Sense Disambiguation with Automatically Acquired Knowledge", 2012.**

In this paper the authors has talked about word sense disambiguation as a process to determining the meaning of the word in context and the limitation in existing methods. In this paper the presented a fully automatic approach for word sense disambiguation and he used two readily available knowledge sources, the first one dictionary and the second one knowledge extracted from unannotated text. The authors also evaluated their method results with Senseval-2 and SemEval 2007, and both large scale WSD evaluation corpora, and require the disambiguation for all words. The authors in their system achieved good performance, and considered their method as better solution to the problem of WSD.

13. **Pankaj Kumar1, Atul Vishwakarma2 and Ashwani Kr. Verma3 Assistant Professor, Computer Science dept., Shri Ramswaroop Memorial Group of Professional Colleges, Lucknow, "Approaches for Disambiguation in Hindi Language "UP, India,, 2013.**

Since the main language of India is Hindi and it's ranking as 4th language among the main languages in the world, but it's still has many ambiguities found in it. The authors in this paper tried to find some solutions for these problems. The authors explained verities methods for disambiguation in this Language. Also they suggested approach to get translation words from dictionary by using a thesaurus to extend query keyword gather. The authors have evaluated the experiments and they showed their suggested method how it has succeeded, because their method successfully can be resolve many works. Also the authors said it will easy to make mix knowledge between Hindi and other Languages.

14. **Rachel Chasin S.B., "Word Sense Disambiguation in Clinical Text", Massachusetts Institute of Technology", June 2013.**

The author in this thesis compared the application of a sem-superviesd public scope state of the art WSI approach in word sense disambiguation in Clinical Text problem. He showed that his approach in clinical domain worked successfully. He also achieved enhancement to the general approach by adding some features in addition bag – of – words it was useful in the disambiguation operations. The author has done many experiments in this thesis and achieved good results in this field. Also he continued to develop and discover new approach to integrate knowledge in his topic as a step for future work.

15. **S.K.Jayanthi and S. Prema, Member, IACSIT, "Word Sense Disambiguation in Web Content Mining Using Brill's Tagger Technique", 2011.**

In this paper the authors have provided detailed discussion in the field of information retrieval (IR) in Web mining. They found in their experiments work that the size of query is a necessary factor in the relationship between the information retrieval and ambiguity. By using Brill's tagger in Web mining, also they tested and analysed word sense disambiguation using for some information retrieval search engines such as Yahoo, Google, msn search, using Brill's tagger. The authors have recorded their result of the experiments in field retrieval by using the disambiguater and made comparison between the two approaches for providing the sense in an information retrieval system; they found that the ambiguity problem useful for retrievals and if the disambigatior is a accurate the disambiguation will be benefit for an information system.

16. **Thanh Phong Pham**, **Hwee Tou Ng** and **Wee Sun Lee, "Word Sense Disambiguation with Semi-Supervised Learning", 2005.**

In this paper the authors has investigated the use of unlabeled training data for WSD, in the framework of semi-supervised learning. Four semi-supervised learning algorithms have been evaluated on 29 nouns of SE2 English lexical sample task and 402 words of SE2 English all-words task. Also they achieved accuracy enhancement from 0.3% to 28.6%, for SE2 English lexical sample task, and about 5 nouns haven't any change in accuracy, and 6 nouns reduced the accuracy ranging from 0.3% to 8%. The empirical results shown that unlabeled data can bring significant improvement in WSD accuracy.

17. **Zhongjian Wang , "Performance Evaluation for Strategy Based on Auto-Adapting Users in Cross Language Information Retrieval", Harbin University of Commerce, Harbin 150028, China, 2013.**

In this paper the author has talked about the relationship between the query and its retrieval, and they said that from verities queries can gain varieties results, depending on viewpoint of the users and their demand. The author in their approach proposed used a thesaurus to extend query keyword gather, and also to compiler words they used dictionary for that. They filtered their study results by using shared words discovered from the correct and incorrect results. He proved that their experiments approach is valid, when they evaluation by inputting Japanese text and outputting Chinese text.

III. CONCLUSION

The literature on word sense disambiguation (WSD) and information retrieval (IR) is too large and difficult to get all the important researchers in this field. The relationship between WSD and IR are still one of the most challenge open problems in the internet. Although search by keywords is the most efficient and popular method to find related information in the internet, it exists two problems by using this relation The first is that some search results don't match the user's requirement. The other is that there are too many similar articles in the search results. Because of the two problems, users spend a lot of time organizing the search results and finding what they really want. Therefore the field of WSD and IR are growing fast, and wishing we can make some contributions in this field.

ACKNOWLEDGMENT

Our thanks to my research guide Dr. Shashank Joshi (Professor at Bharati Vidyapeeth University, College of Engineering), who provided very useful feedback on the first stage of this literature. Also, I would like to thank my parents for their deep love and their support to pursue my goals.

REFERENCES

[1] Antonio Jimeno-Yepes, Alan R. Aronson, National Library of Medicine, 8600 Rockville Pike, Bethesda, 20894, MD, USA, "Self-training and co-training in biomedical word sense disambiguation", 2011.

[2] Anna Bryniarska, "The Paradox of the Fuzzy Disambiguation in the Information Retrieval", Electrical and Computer Engineering, Opole University of Technology, Opole, Poland. 2013.

[3] Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, and Aijun An, "Cross-Lingual Word Sense Disambiguation for Languages with Scarce Resources" Department of Computer Science and Engineering, York University, Canada, 2013.

[4] Caden Howell, Information Retrieval, CSC 575, Dr. Joe Phillips, "Information Retrieval Database with WordNet Word Sense Disambiguation", 2009.

[5] Christopher M. Stokoe, Prof. John Tait, "Automated Word Sense Disambiguation for Internet Information Retrieval", 2002.

[6] Christopher Stokoe, Michael P. Oakes, John Tait, The University of Sunderland Informatics Centre, "Word Sense Disambiguation in Information Retrieval Revisited", 2003.

[7] Francis de la C. Fernández REYES, Exiquio C. Pérez LEYVA, Rogelio Lau FERNÁNDEZ Instituto Superior Politécnico, José Antonio Echeverría, Marianao, Cuba, "Word Sense Disambiguation in Information Retrieval", 2009.

[8] Hwee Tou Ng, "Does Word Sense Disambiguation Improve Information Retrieval?" Department of Computer Science National University of Singapore", 2012.

[9] Mark Sanderson, "word sense disambiguation and information retrieval", department of computing science, university of Glasgow, United Kingdom, 1995.

[10] NITIN INDURKHYA FRED J. DAMERAU, "HANDBOOK OF NATURAL LANGUAGE PROCESSING SECOND EDITION", by Taylor and Francis Group, LLC, 2010.

[11] Pushpak Bhattacharyya, Mitesh Khapra, Indian Institute of Technology Bombay, India, Word Sense Disambiguation, Chapter 2, 2010.

[12] Ping Chen and Chris Bowes, University of Houston-Downtown, Wei Ding and Max Choly, University of Massachusetts, Boston, "Word Sense Disambiguation with Automatically Acquired Knowledge", 2012.

[13] Pankaj Kumar1, Atul Vishwakarma2 and Ashwani Kr. Verma3 Assistant Professor, Computer Science dept., Shri Ramswaroop Memorial Group of Professional Colleges, Lucknow, "Approaches for Disambiguation in Hindi Language "UP, India,, 2013.

[14] Rachel Chasin S.B., "Word Sense Disambiguation in Clinical Text", Massachusetts Institute of Technology", June 2013.

[15] S.K.Jayanthi and S. Prema, Member, IACSIT, "Word Sense Disambiguation in Web Content Mining Using Brill's Tagger Technique", 2011.

[16] Thanh Phong Pham, Hwee Tou Ng and Wee Sun Lee, "Word Sense Disambiguation with Semi-Supervised Learning", 2005.

[17] Zhongjian Wang , "Performance Evaluation for Strategy Based on Auto-Adapting Users in Cross Language Information Retrieval", Harbin University of Commerce, Harbin 150028, China, 2013.

First Author: Boshra F. Zopon AL_Bayaty received her B.E degree in computer science from AL_Mustansiriyah University, College of Education in 2002. And received her M.S.C degree in computer science from Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies. Doing her the PH.D. Computer Science at Bharati Vidyapeeth Deemed University, Pune. She is currently working in the Ministry of Higher Education & Scientific Research, AL_Mustansiriyah University in Iraq/ Baghdad. Her research interests include software engineering.

Second Author: Shashank Joshi received his B.E. degree in Electronics and Telecommunication from Govt. College of Engineering, Pune in 1988, the M.E. and Ph. D. Degree in Computer Engineering from Bharati Vidyapeeth Deemed University Pune. He is currently working as the Professor in Computer Engineering Department Bharati Vidyapeeth Deemed University College of Engineering, Pune. His research interests include software engineering. Presently he is engaged in SDLC and secure software development methodologies. He is innovative teacher devoted to Education and Learning for the last 23 yrs.