Journal of Software Engineering and Applications, 2012, 5, 175-180
doi:10.4236/jsea.2012.512b034 Published Online December 2012 (http://www.SciRP.org/journal/jsea)

Using Wikipedia as an External Knowledge Source for Supporting Contextual Disambiguation

Shahida Jabeen, Xiaoying Gao, Peter Andreae

School of Engineering and Computer Science, Victoria University of Wellington, New Zealand.
Email: shahidarao@ecs.vuw.ac.nz, Xiaoying.Gao@ecs.vuw.ac.nz, Peter.Andreae@ecs.vuw.ac.nz

Received 2012

ABSTRACT

Every term has a meaning but there are terms which have multiple meanings. Identifying the correct meaning of a term in a specific context is the goal of Word Sense Disambiguation (WSD) applications. Identifying the correct sense of a term given a limited context is even harder. This research aims at solving the problem of identifying the correct sense of a term given only one term as its context. The main focus of this research is on using Wikipedia as the external knowledge source to decipher the true meaning of each term using a single term as the context. We experimented with the semantically rich Wikipedia senses and hyperlinks for context disambiguation. We also analyzed the effect of sense filtering on context extraction and found it quite effective for contextual disambiguation. Results have shown that disambiguation with filtering works quite well on manually disambiguated dataset with the performance accuracy of 86%.

Keywords: Contextual Disambiguation; Wikipedia Hyperlinks; Semantic Relatedness

1. Introduction

Ambiguity is implicit to natural languages with a large number of terms having multiple contexts. For instance, the English noun “bank" can be a financial institution or it could mean river side. Polysemy is the capability of a term to have multiple meanings and a term with multiple senses is termed as “polysemous term”. Resolving ambiguity and identifying the most appropriate sense is a critical issue in natural language interpretation and processing. Identification of the correct sense for an ambiguous term requires understanding of the context in which it occurs in natural language text. Word Sense Disambiguation is defined as the task of automatically selecting the most appropriate sense of a term within a given context from a set of available senses [1-4].

What is the common context of Bank and flood? What sense of Tree comes to mind when it occurs with computer? In both cases, the relationship of each pair of terms is strong but can’t be measured unless considering the correct senses of either or both terms. The first pair of terms has a cause and effect relation provided the correct context of bank i.e Levee is taken into account. Whereas, to relate the other term pair, the term Tree should be mapped to Tree (data structure). Making judgments about deciphering the relevant context of a term is an apparently ordinary task but actually requires a vast amount of background knowledge about the concepts.

The task of contextual information extraction for disambiguation of natural language text relies on knowledge from a broad range of real world concepts and implicit relations [5]. To effectively perform such a task, computers may require an external knowledge source to infer implicit relations. External knowledge sources can vary from domain specific thesauri to lexical dictionaries and from hand crafted taxonomies to knowledge bases. Any external knowledge source should have the following three characteristics to sufficiently support contextual disambiguation:

- **Coverage:** The coverage of the external knowledge source is the degree to which it represents collected knowledge. It should be broad enough to provide information about all relevant concepts.
- **Quality:** Information provided by the external knowledge source should be accurate, authentic and up-to-date.
- **Lexical semantic resource:** The knowledge base should encode rich lexical and structural semantics.

Fortunately, Wikipedia, a web based freely available encyclopedia, comes up with all the necessary characteristics to support contextual information extraction and disambiguation. As a knowledge source, Wikipedia not only focuses on general vocabulary but also covers a large number of named entities and domain specific terms. The specialty of Wikipedia is that each of its articles is dedicated to a single topic with an additional benefit of heavy linking between articles. Wikipedia also covers specific senses of a term, surface forms, spelling variations, abbreviations and derivations. Importantly, it has a semantically rich hyperlink network relating articles that cover different types of lexical semantic relations such as hyponymy and hypernymy, synonymy, antonymy and polysemy. Hence, Wikipedia sufficiently demonstrates all the necessary aspects of a good knowledge source. It provides a knowledge base for extracting information in a more structured way than a search engine and with a coverage better than other knowledge sources [6].

The main goal of this research is to use Wikipedia as the external knowledge source to decipher true meaning of a term using the other term as the context. In particular, we have taken into account Wikipedia hyperlinks structure and senses for disambiguating the context of a term pair. The rest of the paper is organized as follows. Section 2 discusses two broad categories of disambiguation work proposed in literature and the corresponding research done in each category. Section 3 discusses our proposed approach for context extraction and disambiguation. Section 4 comprises of the performance analysis of our proposed methodology using a manually designed dataset consisting of English term pairs. Finally, section 5 concludes our research and discusses some future research directions.

2. Related Work

Word Sense Disambiguation (WSD) is a well explored area with lot of research work still going on in context identification and disambiguation. Research in this area can be broadly categorized into two main streams [7]: Lexical approaches and taxonomic approaches.

Lexical approaches are based either on the analysis of a disambiguated corpus or by extracting strings from the definition of that sense. Such approaches are based on identifying the lexical features such as occurrence statistics or co-occurrence computations. By contrast, taxonomic approaches identify correct nodes in the hierarchy of senses and explore the relations between nodes for computing the semantic closeness.

Early efforts in Lexical approaches were based on machine readable dictionaries and thesauri, associating word senses with short definitions, examples or lexical relations. A simple approach of this type is based on comparing dictionary definitions of words, also called glosses, to the context words (the words appearing in the surrounding text) of an ambiguous word. Clearly, the higher the overlap between context words and the dictionary definitions of a particular sense of the ambiguous word, the higher the chances of getting the correct sense for the word. Cowie et al. [8] and Lesk [9] based their approaches on machine readable dictionaries and thesauri. Pederson et al. [10] adapted the Lesk algorithm for word sense disambiguation by using lexical database WordNet instead of standard dictionary as a source of glosses. They exploited the hierarchy of WordNet semantic relations for disambiguation task. Patwardhan et al. [1] generalized adapted Lesk algorithm to a method of term sense disambiguation. They used the gloss overlap as an effective measure of semantic relatedness. Pedersen [11] further explored the use of similarity measures based on path findings in concept networks, information contents derived from large corpus and term sense glosses. They concluded that the gloss based measures were quite effective for term sense disambiguation. Yarowsky [12] used Statistical models on Roget's thesaurus categories to build context discriminators for the word senses that are members of conceptual classes. A conceptual class such as "machine" or "animal" tends to appear in recognizably different contexts. They also used the context indicators of Roget's thesaurus as the context indicators for the members of conceptual categories.

Following taxonomic approaches, Agirre [13] proposed a method of lexical disambiguation over Brown's Corpus using noun taxonomy of WordNet. He computed the conceptual density by finding the combination of senses from a set of contiguous nouns that maximized conceptual distances in taxonomic concepts. Veronis et al. [14] automatically build a very large Neural Network from definition text in machine readable dictionaries and used this network for word sense disambiguation. They used the node activation scheme for moving closer to the most related senses following the "Winner-take-all" strategy in which every active node sent an activation to another increasingly related node in the network. Mihalcea et al. [15] used the page ranking algorithm on Semantic Networks for sense disambiguation. Iterative graph-based ranking algorithms are essentially a way of deciding the importance of a node within a graph. When one node in a graph is connected to another one, it is casting a vote for that other node. The higher the number of votes of a node , the higher the importance of that node. They find the sysnet with highest PageRank score for each ambiguous word in the text assuming that it will uniquely identify the sense of the word.

Wikipedia and Sense Disambiguation

Availability of free online thesaurus, dictionaries and encyclopedias and other knowledge sources has scaffolded the improvement in both lexical and taxonomic based word sense disambiguation. Different methodologies are found in literature for computing term sense disambiguation based on Wikipedia as the external knowledge source. Ponzetto and Navigli [4] addressed the problem of knowledge acquisition in term sense disambiguation. They proposed a methodology for extending WordNet with large amount of semantic relations derived from Wikipedia. They associated Wikipedia pages with the WordNet senses to produce a richer lexical resource. Rada [2] used Wikipedia as a source of sense annotation for generating sense-tagged data for building accurate and robust sense classifiers. Turdakov and Velikhov [16] proposed a semantic relatedness measure based on Wikipedia links and used it to disambiguate terms. They proposed four link-based heuristics for reducing the search space of potentially related topics. Fogarolli [17] used Wikipedia as a reference to obtain lexicographic relations and combined them with the statistical information to deduce concepts related to terms extracted from a corpus. Cucerzan [18] presented an approach for the recognition and semantic disambiguation of named entities based on agreement between information extracted from Wikipedia and the context of Web search results. Bunescu [19] addressed the same problem of detecting and disambiguating named entities in open domain text using Wikipedia as the external knowledge source.

The kind of problem that we address in this research is a variant of the main stream of term sense disambiguation research, where the aim is to identify the context of a single term. We look to find out the context of two terms with respect to each other. This task is critical in many approaches involving relatedness computation [6,20, 21].

3. Disambiguation Methods

There are two approaches that we have adopted for contextual extraction and disambiguation along with their variants based on two factors: relatedness measure and sense filtering.

3.1. WikiSim Based Disambiguation

Our main approach for contextual disambiguation using Wikipedia consists of three phases: Context Extraction, in which we extract the candidate context of both input terms; Context Filtering, where we filter out certain contexts based on their type, thus avoiding unnecessary context; Contextual Disambiguation, where semantic similarity between candidate contexts is computed using Wikipedia hyperlink based

1) Context Extraction: We start with identification of all possible contexts corresponding to each input term based on Wikipedia senses. Each Wikipedia article is associated with a number of senses. For instance, there are various senses for terms Present and Tense. The best sense of Present would be Present tense and the best sense of tense would be Grammatical tense in correct context of each other. The aim of this phase is to extract all possible contexts corresponding to the input term pair. For this purpose, we used Wikipedia disambiguation pages to extract all listed senses as candidate senses and populate them in the context set of each input term.

2) Context Filtering: There are three broad categories of manually annotated Wikipedia senses.

- Senses with parenthesis
- Single term senses
- Phrasal senses

The first type of senses are those Wikipedia senses which are generic and include broader context within parenthesis following the title. For example, Crane (Bird) and Crane (machine) are two different contexts of the term Crane. Both of these senses are quite distinctive and cover broader contexts of machine and bird.

In the second category of senses, single term context falls. These contexts cover certain very important lexical relations such as synonymy, hypernymy and hyponymy and derivations. For instance, Gift is a synonym of the term present, similarly carnivores is the hypernym of the term tiger. It is found that such senses are very useful for contextual information extraction.

The third type of Wikipedia senses are phrasal senses, which usually have very specific and limited context. This context can have characteristics of the more general sense but it would be focused more on some other specific features which cannot be considered true for the generalized sense. For instance, corresponding to a term forest, the phrasal sense might be Forest Township, Missaukee County, Michigan which discusses a geographically specific context rather than the more general context of forest. Such type of senses might not be very useful in extracting the contextual information. For this reason we excluded this third type of senses from the context set of each input term. Experiments proved that bigram senses still contain the general context of a term. So, we gathered all uni-gram and bi-gram senses, senses with parenthesis and senses shared by both input terms and put them in the context set corresponding to each input term.

3) Contextual Disambiguation: In contextual disambiguation, the first step is to extract all unique inlinks (all articles referring to the input term article) and outlinks (all articles referred by the input term article) corresponding to each candidate sense of the context set. The link vector of each candidate sense of first input term is compared with that of each sense of the second input term. The assumption behind this comparison is to find out those senses which share maximum number of links, thus indicating a strong relatedness. Each sense pair is assigned a weight based on a relatedness measure called WikiSim [20], as shown below:

$$rel(s_i, s_j) = \left[\frac{|S|}{|T|}\right] \times 2 \qquad (1)$$

In the above formula $s_i \in | Sw_a |$, where $|Sw_a |$ is the set of all senses of term w_a whereas, s_j sense corresponds to input term w_b and $s_i \in |\ Sw_b|$. S is the set of all the links shared between a sense pair and T represents total number of links of both senses. In other terms, the weight of a sense pair is the link probability of shared links, or is 0 if shared links do not exist. Once we get scores for all sense pairs, the next step is to find out the sense pair with highest score, thus getting most closely related senses of both input terms. For this purpose, all sense pairs are sorted based on their WikiSim score and the sense pair having the highest weight is taken as the disambiguated context corresponding to the input term pair.

3.2. WLM based Disambiguation

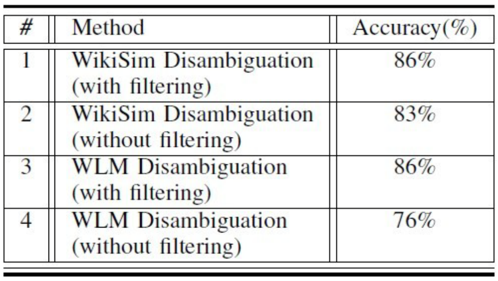
In order to evaluate the performance of our main approach and to analyze the effect of using a different relatedness measure on disambiguation task, we used WLM relatedness measure [21] which is also based on Wikipedia hyperlinks and is proven to be quite effective in computing term relatedness. We followed the same methodology as the WikiSim based Disambiguation for extracting candidate senses and populating the context vector but replaced the WikiSim relatedness measure with WLM measure while computing the sense pair scores.

4. Evaluation

4.1. Experimental setup

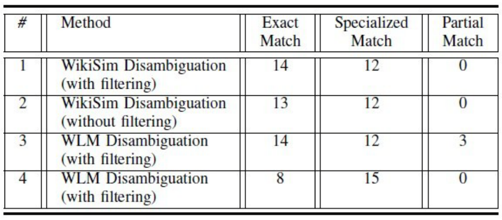
We used the version of Wikipedia released in July, 20111. At this point, it contains 31GB of uncompressed XML markup which corresponds to more than five million articles which sufficiently covers all the concepts for which manual judgment were available. To explore and easily draw upon the contents of Wikipedia, we used the latest version (wikipedia-miner-1.2.0) of Wikiminer toolkit [22] which is an open source Java code2. Since the problem addressed in this research is a variant of the standard disambiguation task, where rather than resolving the context of a single term we do that for a pair of terms considering each word as the context for the other word, we needed a different dataset of disambiguated term pairs. So, we used a manually designed dataset named WikiContext as shown in Table 3. It consists of 30 English term pairs which are manually disambiguated to corresponding Wikipedia articles in context of the other input term.

Table 1. Accuracy Based Performance Comparison Of Disambiguation Methods.



#	Method	Accuracy(%)
1	WikiSim Disambiguation (with filtering)	86%
2	WikiSim Disambiguation (without filtering)	83%
3	WLM Disambiguation (with filtering)	86%
4	WLM Disambiguation (without filtering)	76%

Table 2. Statistics Of Types Of Disambiguation Performed By Each Method.



#	Method	Exact Match	Specialized Match	Partial Match
1	WikiSim Disambiguation (with filtering)	14	12	0
2	WikiSim Disambiguation (without filtering)	13	12	0
3	WLM Disambiguation (with filtering)	14	12	3
4	WLM Disambiguation (with filtering)	8	15	0

4.2. Experimental Results and Discussion

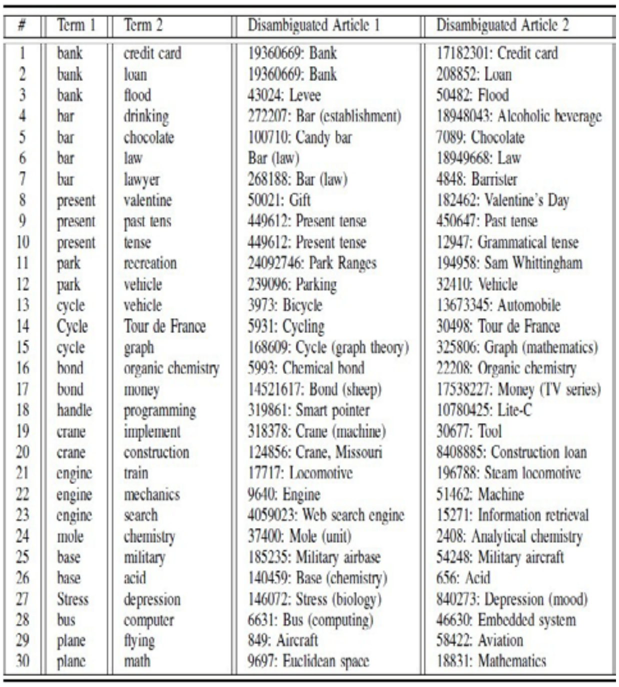
To compare performance of our proposed methods, we automatically disambiguated term pairs in the dataset and compared them with the manually disambiguated Wikipedia articles. To measure the performance of each method, we used the accuracy measurement:

thod defined as follow:

$$Accuracy(method) = \frac{count(|P_c|)}{count(|P_t|)} \times 100 \qquad (2)$$

where, $|P_c|$ is the set of correctly disambiguated term pairs and $|P_t|$ refers to the set of all disambiguated term pairs. In other words, it is the ratio of correct disambiguation and the size of the dataset.

Table 3. Wikisim (With Filtering):Word Pairs And Corresponding Automatic Disambiguation.



#	Term 1	Term 2	Disambiguated Article 1	Disambiguated Article 2
1	bank	credit card	19360669: Bank	17182301: Credit card
2	bank	loan	19360669: Bank	208852: Loan
3	bank	flood	43024: Levee	50482: Flood
4	bar	drinking	272207: Bar (establishment)	18948043: Alcoholic beverage
5	bar	chocolate	100710: Candy bar	7089: Chocolate
6	bar	law	Bar (law)	18949668: Law
7	bar	lawyer	268188: Bar (law)	4848: Barrister
8	present	valentine	50021: Gift	182462: Valentine's Day
9	present	past tens	449612: Present tense	450647: Past tense
10	present	tense	449612: Present tense	12947: Grammatical tense
11	park	recreation	24092746: Park Ranges	194958: Sam Whittingham
12	park	vehicle	239096: Parking	32410: Vehicle
13	cycle	vehicle	3973: Bicycle	13673345: Automobile
14	Cycle	Tour de France	5931: Cycling	30498: Tour de France
15	cycle	graph	168609: Cycle (graph theory)	325806: Graph (mathematics)
16	bond	organic chemistry	5993: Chemical bond	22208: Organic chemistry
17	bond	money	14521617: Bond (sheep)	17538227: Money (TV series)
18	handle	programming	319861: Smart pointer	10780425: Lite-C
19	crane	implement	318378: Crane (machine)	30677: Tool
20	crane	construction	124856: Crane, Missouri	8408885: Construction loan
21	engine	train	17717: Locomotive	196788: Steam locomotive
22	engine	mechanics	9640: Engine	51462: Machine
23	engine	search	4059023: Web search engine	15271: Information retrieval
24	mole	chemistry	37400: Mole (unit)	2408: Analytical chemistry
25	base	military	185235: Military airbase	54248: Military aircraft
26	base	acid	140459: Base (chemistry)	656: Acid
27	Stress	depression	146072: Stress (biology)	840273: Depression (mood)
28	bus	computer	6631: Bus (computing)	46630: Embedded system
29	plane	flying	849: Aircraft	58422: Aviation
30	plane	math	9697: Euclidean space	18831: Mathematics

When compared performance accuracy of both WikiSim based disambiguation and WLM based disambiguation, WikiSim based disambiguation is found to be comparable to that of WLM based disambiguation, both having an accuracy of 86% as shown in Table I. To analyze the effect of applying sense filtering in both disambiguation approaches, we skipped context filtering step from each approach and performed disambiguation with all possible contexts. Detailed analysis of each approach is summarized in Table II. Three types of disambiguation are taken into account in this research: First, those word pairs which are matched exactly to the manual disambiguation; second, those word pairs which are matched to a specialized area or subtopic of the correct context; third, those word pairs in which one of the term is correctly disambiguated in context of the other word but the other term is disambiguated to a wrong context. Results of WikiSim based disambiguation revealed that there was no partial match in case of both context filtering and without filtering. Whereas, some of the specialized matches were found to be more relevant then the exact matches. For example in case of the term pair <mole, chemistry>, mole was disambiguated exactly to Mole (unit) which is the measurement of amount of chemical substance but chemistry was disambiguated to even more related context Analytical Chemistry which deals with quantification of the chemical components. Similarly, the term pair <bar, drinking> was disambiguated to <Bar (Establishment), Alcoholic beverages>. In case of WLM based disambiguation with filtering, three term pairs were found to be partially matched to the correct context. Table III shows the results on WikiSim based disambiguation (with filtering) on the dataset WikiContext. It shows disambiguated Wikipedia articles corresponding to input term pairs.

Table I
Accuracy Based Performance Comparison Of Disambiguation Methods

Overall, disambiguation based on filtering is found to be better than the one without filtering. The accuracy of WikiSim based method is 3% increased when filtering is applied. The effect of filtering became more evident in WLM where the accuracy of filtering based disambiguation increased to 10%. The results of our main approach are quite encouraging and comparable to the WLM based disambiguation. In order to avoid any bias in the results due to smaller size of dataset and to test the effectiveness of our approach more critically, we plan to use some bigger dataset in future. To the best of our knowledge, there is no dataset available that addresses this particular kind of problem. One limitation of our approach is that it relies heavily on Wikipedia senses, which are manually encoded and may not sufficiently cover all possible contexts of some terms and suffers from inconsistent formatting due to manual encoding. We believe that using other Wikipedia features such as anchor texts, categories, hyperlinks and redirects for semantic extraction would definitely help in this regard.

5. Conclusion

In this paper we proposed and evaluated a novel approach for extracting contextual information from Wikipedia and using it to disambiguate a term using a single term as a given context. Based on Wikipedia hyperlink structure and senses, our approach used WikiSim, a Wikipedia based relatedness measure to compute scores of sense pairs and compare them based on their relatedness. For the sense disambiguation, we extracted various senses of first input term and disambiguated each sense in various contexts of the other input term. We evaluated the performance of our approach by comparing it with WLM based disambiguation approach and analyzed the effect of context filtering disambiguation. Results have shown that with an accuracy of 86%, our approach performs quite well when compared with manually disambiguated dataset of term pairs. In future, we plan to apply this disambiguation approach along with the semantic relatedness in key phrase clustering task for an indirect evaluation of our approach on a bigger dataset to avoid any bias in the current results due to smaller dataset size.

REFERENCES

[1] S. Patwardhan, S. Banerjee, and T. Pedersen, “Using measures of semantic relatedness for word sense disambiguation,” in Proceedings of the 4th International Conference on IntelligentText Processing and Computational Linguistics, February 2003, pp. 241–257.

[2] R. Mihalcea, “Using wikipedia for automatic word sense disambiguation,” in North American Chapter of the Association for Computational Linguistics (NAACL 2007), 2007.

[3] D. McCarthy, “Word sense disambiguation: The case for combinations of knowledge sources,” Natural Language Engineering, vol. 10, pp. 196–200, June 2004.

[4] S. P. Ponzetto and R. Navigli, “Knowledge-rich word sense disambiguation rivaling supervised systems,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1522–1531.

[5] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, “Wikiwalk: random walks on wikipedia for semantic relatedness,” in 2009 Workshop on Graph-based Methods for Natural Language Processing, 2009, pp. 41–49.

[6] M. Strube and S. P. Ponzetto, “Wikirelate! computing semantic relatedness using wikipedia,” in proceedings of the 21st national conference on Artificial intelligence, vol.

2, 2006, pp. 1419–1424.

[7] J. Curtis, J. Cabral, and D. Baxter, "On the application of the cyc ontology to word sense disambiguation," in Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference, 2006, pp. 652–657.

[8] J. Cowie, J. Guthrie, and L. Guthrie, "Lexical disambiguation using simulated annealing," in Proceedings of the workshop on Speech and Natural Language, 1992, pp. 238–242.

[9] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in Proceedings of the 5th annual international conference on Systems documentation, 1986, pp. 24–26.

[10] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," In Proceeing of the Third International Conference on Intelligent Text Processing and Computational Linguistics, 2002, pp. 136–145.

[11] T. Pedersen, S. Banerjee, and S. Patwardhan, "Maximizing Semantic Relatedness to Perform Word Sense Disambiguation," University of Minnesota Supercomputing Institute, Research Report UMSI 2005/25, March 2005.

[12] D. Yarowsky, "Word-sense disambiguation using statistical models of roget's categories trained on large corpora," in Proceedings of the 14th conference on Computational linguistics - Volume 2, 1992, pp. 454–460.

[13] E. Agirre and G. Rigau, "Word sense disambiguation using conceptual density," in Proceedings of the 16th conference on Computational linguistics - Volume 1, 1996, pp. 16–22.

[14] J. Veronis and N. M. Ide, "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries," in Proceedings of the 13th conference on Computational linguistics - Volume 2, 1990, pp. 389–394.

[15] R. Mihalcea, P. Tarau, and E. Figa, "Pagerank on semantic networks, with application to word sense disambiguation," in Proceedings of the 20th international conference on Computational Linguistics, 2004.

[16] D. Turdakov and P. Velikhov, "Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation," SYRCoDIS, vol. 355, pp. 1–6, 2008.

[17] A. Fogarolli, "Word sense disambiguation based on wikipedia link structure," in Proceedings of the 2009 IEEE International Conference on Semantic Computing, 2009, pp. 77–82.

[18] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in Proceedings of EMNLP-CoNLL 2007, 2007, pp. 708–716.

[19] B. Razvan and P. Marius, "Using encyclopedic knowledge for named entity disambiguation," in Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 2006, pp. 9–16.

[20] S. Jabeen, X. Gao, and P. Andreae, "Improving contextual relatedness computation by leveraging wikipedia semantics," in 12th Pacific Rim International Conference on Artificial Intelligence (To appear), 2012.

[21] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, 2008, pp. 25–30.

[22] D. Milne, "An open-source toolkit for mining Wikipedia," in Proceeding of New Zealand Computer Science Research Student Conference, vol. 9, 2009.