Graph Algorithms for Word Sense Disambiguation in Biomedicine

Rodrigo Rafael Villarreal Goulart[1], Juliano Varella de Carvalho[1], Vera Lucia Strube de Lima[2]

[1] ICET – Universidade Feevale – Novo Hamburgo, RS – Brazil
[2] FACIN – PPGCC – PUCRS – Porto Alegre, RS – Brazil

{rodrigo,julianovc}@feevale.br, vera.strube@pucrs.br

Abstract. *Word Sense Disambiguation (WSD) is an important task for Biomedicine text-mining. Supervised WSD methods have the best results but they are complex and their cost for testing is too high. This work presents an experiment on WSD using graph-based approaches (unsupervised methods). Three algorithms were tested and compared to the state of the art. Results indicate that similar performance could be reached with different levels of complexity, what may point to a new approach to this problem.*

1. Introduction

The process of selecting the accurate sense of a word is called *Word Sense Disambiguation* (WSD). Identifying the correct sense of words helps to improve text-mining systems. Among the approaches developed, there are those with specific purpose, in which WSD is restricted to a particular knowledge domain. In the biomedical domain, knowledge-rich systems are projected employing Natural Language Processing (NLP) techniques to deal with the ambiguity inherent to texts. MedLEE[1] and PubMed[2] are examples of such systems. MedLEE extracts information from radiology texts. It organizes and classifies information like a controlled vocabulary. PubMed is an indexer of biomedical articles. In both cases, the search for information is associated with the identification and classification of concepts present in texts.

However, the process of automatically identifying the accurate sense of a word in a text is still an open problem. For example, consider the search for the word *glucose* in the PubMed indexer. According to the UMLS (Unified Medical Language System) metathesaurus (Humphreys et al. 1998), specialized in Biomedicine, the word *glucose* is present in three concepts: *glucose*, *plasma glucose measurement* and *glucose measurement*. The user searching for the word *glucose* in the PubMed indexer might be unaware of, or even not desire, the results with the *plasma glucose measurement* and *glucose measurement* concepts. To identify the accurate sense of a word, the context in which it was used plays a very important role. Generally, concepts or simply the surrounding words (i.e. words that are before and after the ambiguous word in a text) represent the context. Together with this kind of information, it is possible to make use of automatic methods that consider the situation in which the word was employed, and then select the most adequate sense according to a predetermined set of possible senses, such as, for example, those established in the UMLS.

1 http://www.cat.columbia.edu/?page_id=84 Last access: 9th April 2015.

2 http://www.ncbi.nlm.nih.gov/pubmed/ Last access: 9th April 2015.

Approaches based on supervised learning have the best results in WSD of Biomedicine texts (Navigli 2012; Preiss and Stevenson 2013; Trivedi et al. 2014). However, they demand labeled examples for training, which might not be available or be too costly to be developed. This limitation means that the supervised approaches may disambiguate a sample of words to which a set of training data was elaborated, and this limits their practical use. On the other hand, unsupervised approaches do not require labeled examples. As they use structured knowledge sources, there is no need for training and testing sets. Unsupervised and semi-supervised approaches have been previously explored with the use of UMLS (Garla and Brandt 2012; Navigli 2012; McInnes and Pedersen 2013). Furthermore, knowledge sources might also be taken as a graph, where the topology can suit the unsupervised method. UMLS, as well as WordNet (Miller 1995), are examples of this case: semantic relations are established between the concepts in form of a graph.

There are specific algorithms (also known as *metrics*) that take into account the structure of a graph and determine the importance of a vertex. The most popular graph-based algorithms associated to information retrieval on the Internet are PageRank (Page et al. 1998) and HITS (Kleinberg 1999). In WSD, the personalized *PageRank*, *Degree Centrality* and *Key Player Problem* algorithms have also been explored in specialized and non-specialized domains (Agirre and Soroa 2009; Agirre et al. 2010; Navigli and Lapata 2010). The results obtained with these algorithms made it possible to identify the best ones for different scenario settings (i.e. knowledge domain, knowledge source, corpora for testing), but there are gaps to be filled.

In such context, the present work proposes a study of the problems and solutions related to WSD in the Biomedicine domain. Three algorithms based on graphs were investigated aiming to compare, identify gaps and broaden the results found until that point. The results indicate that those algorithms employed for general knowledge domain do not behave in the same way in the specific domain of Biomedicine. Furthermore, the experiment reveals that results with similar performances may be reached with algorithms of different complexity levels, what might bring gain on computational processing time.

2. Word Sense Disambiguation

WSD using lexical samples each ambiguous word is associated to a list of possible meanings, called candidates, usually related to some dictionary. Information about the ambiguous word is then used to select its sense. This information may be comprised in the text itself or in external sources. The information comprised in the text can be, for example, the surrounding words of the ambiguous word (context) or the morphosyntactic categories of these words. External sources contain additional information about the ambiguous word, its context or the candidates for sense disambiguation. Annotated corpora, ontologies and thesauri are examples of external knowledge sources.

2.1. Graph Approaches for WSD

Structured sources are one type of external knowledge source that can be employed in WSD. They depict semantic relations among concepts, whether of specialized domain

or not, which enable automated processing. Thesauri and ontologies are examples of this kind of resource (Navigli 2009).

UMLS is considered a structured resource. It represents the unification of a broad set of controlled vocabularies of Biomedicine, besides classification systems. The organization of this metathesaurus is based on concepts identified by a *Concept Unique Identifier* (CUI). For example, the following CUIs are associated to the term "*Adjustment*": C0376209 (*Individual Adjustment*), C0456081 (*Adjustment Action*) and C0683269 (*Psychological Adjustment*). The metathesaurus also contains information about relations among CUIs arranged in database tables. Using these tables it is possible to build a graph that represents the concepts as nodes and the relations as edges. This approach can be employed in the representation of candidate concepts and in the representation of the ambiguous word context as well. So graph structure can be used to determine the importance of nodes in selecting senses. Therefore, it is necessary to establish a method to build the graph from this information, to this end.

2.2. Building graphs from texts

Navigli and Lapata (2007) present a method to build such a graph. Consider the following paragraph as an example to understand better this process:

> *… and the regression coefficient of percentage decline in FEV1 with log dose, were calculated ("slope", after transformation), with and without calibration of nebulizers by weight and* ***adjustment*** *for nonresponse bias. Standardization for baseline lung function and variation in smoking prevalence was applied to slope. Results were ...*

This text was extracted from the NLM-WSD corpus proposed in (Weeber et al. 2001). The corpus has 50 ambiguous annotated words. For each of these ambiguous words, 100 instances were annotated. A total of 5000 abstracts forms the corpus. The abstracts were randomly extracted from the MEDLINE base, year 1998. The instances were manually disambiguated by 11 annotators who annotated each occurrence of the term with the corresponding meaning in UMLS. The word *adjustment*, for example, has three possible meanings, indicated in the corpus as: *Individual Adjustment*, *Adjustment Action* and *Psychological Adjustment*. In this case, the first option represents the sense chosen by the annotators. Some instances were classified as *none* to indicate that the annotators did not find a possible meaning for the term in UMLS.

Other concepts in the context must be identified to build the *G* graph that represents the terms present in the context of the ambiguous word. Considering a window of 20 concepts, 10 before the ambiguous word and 10 after it, we have the following annotation:

> *… and the regression coefficient of* $[percentage]_{-10}$ *decline in* $[FEV1]_{-9}$ *with* $[log]_{-8}$ $[dose]_{-7}$*, were* $[calculated]_{-6}$ *("*$[slope]_{-5}$*", after* $[transformation]_{-4}$*), with and without* $[calibration]_{-3}$ *of* $[nebulizers]_{-2}$ *by* $[weight]_{-1}$ *and* $\mathbf{[adjustment]_0}$ *for nonresponse* $[bias]_{+1}$*.* $[Standardization]_{+2}$ *for* $[baseline]_{+3}$ $[lung\ function]_{+4}$ *and* $[variation]_{+5}$ *in* $[smokin]_{+6}$ $[prevalence]_{+7}$ *was* $[applied]_{+8}$ *to* $[slope]_{+9}$*.* $[Results]_{+10}$ *were ...*

The words between square brackets determine the concepts and the right brackets are followed by their position in relation to the ambiguous word. For example, the sixth

word before the ambiguous word is *calculated*. Compound words can be found (e.g. $[lung\ function]_{+4}$). As the concepts in the context are those found in the UMLS metathesaurus, they can be ambiguous, such as in the case of $[variation]_{+5}$. In this situation, the first sense found is the one used in this example.

Considering the paragraph previously presented and the three possible meanings for the term *adjustment*, each candidate and its context allow the generation of a different graph. An example is the graph of the candidate sense *Psychological Adjustment* (C0683269), presented in **Figure 1**. In this graph, the CUIs of each concept in the paragraph are expressed according to UMLS. The grey ellipses represent the concepts found in the context. The gray rectangle represents the candidate term for disambiguation. The others are terms that establish an indirect relation between the candidate concept and the ones found in the context.

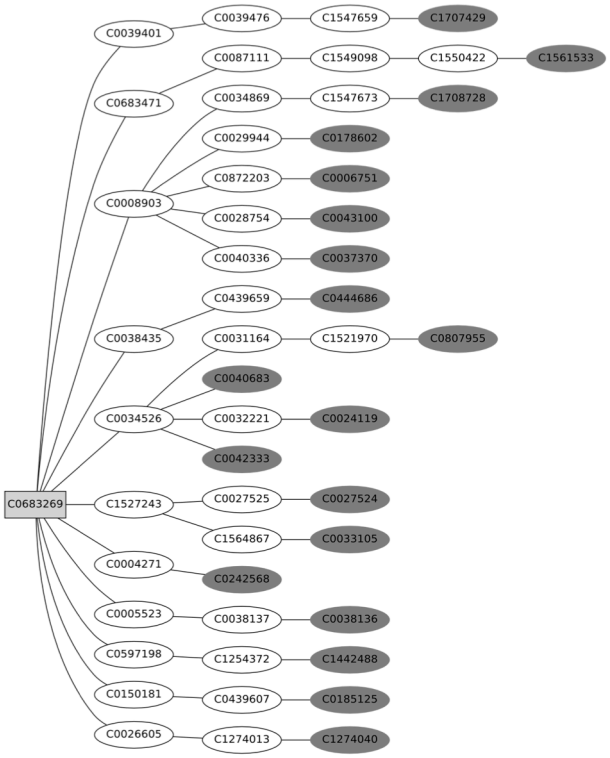

Figure 1. Graph for the *Psychological Adjustment* concept

2.3. Connectivity algorithms

The classification of the candidate vertex is necessary, according to its importance, to select the accurate sense, based on some connectivity degree measure. There are several proposals for this, among which the ones that obtained the best unsupervised results were selected, according to (Navigli and Lapata 2007) and (Navigli and Lapata 2010). Consider E as the set of all edges and V as the set of all vertices.

Degree Centrality, or simply Degree, is the simplest way to measure the importance of a vertex. It is determined by its degree, that is, the number of vertex edges. Thus, we have:

$$\deg(v) = \left|\{\{u,v\} \in E : u,v \in V\}\right| \qquad (1)$$

where the degree of a *v vertex*, *deg(v)*, is determined by the number of edges between *v* and each vertex in the graph. A vertex is central if and only if it has a high degree. In the same way, an unconnected vertex is zero degree. The centrality degree is normalized by its maximum degree, that is, the number of vertices in the graph except itself. Thus, we have:

$$C_D = \frac{\deg(v)}{|V|-1} \tag{2}$$

Within the ***Key Player Problem*** (KPP), a vertex is considered important if and only if it is relatively close to all other vertices (Borgatti 2006). Thus, we have:

$$KPP(v) = \frac{\sum_{u,v \in V: u \neq v} \frac{1}{d(u,v)}}{|V|-1} \tag{3}$$

where the numerator is the sum of the inverses of the distances between *v* and all the other nodes. The denominator is the number of nodes in the graph, except for *v*.

The PageRank algorithm (Page et al. 1998; Brin and Page 1998) is a method for classifying graph vertices according to their relative structural importance. A variation of PageRank employed in WSD is the ***personalized PageRank*** algorithm (Haveliwala 2002). It calculates the structural importance of a graph's vertices when some of them are more relevant than others for a specific situation. Be *M* a transition probability matrix *N x N*, where $M_{ij} = 1/d_i$ (inverse of the degree of an *i* vertex) if there is a way from v_i to v_j, otherwise it is zero. Be *w* a normalized stochastic vector *N x 1* whose values are all 1/*N*. Then, the *P PageRank* Vector over the *G* graph is obtained by the following equation:

$$P = cMP + (1-c)w \tag{4}$$

In PageRank the *w* vector is evenly distributed, thus determining equal probabilities to all vertices in the graph when there are random jumps. However, in the personalized PageRank the *w* vector might be non-uniform and determine higher probabilities for specific vertices, conducting to preferential vertices. Additional examples and details can be found in (Agirre et al. 2010).

Navigli and Lapata (2010) carried on an experimental study applying eight connectivity algorithms over three corpora. The objective was to compare the performance of these algorithms, and to study them against two baselines. Navigli and Lapata used WordNet as a knowledge source (graph) to distinguish the senses, as well as the lexical and semantic relations of the corpora. According to Navigli and Lapata, the experiments' results with the *Degree* and PageRank algorithms are statistically similar. Apart from the external knowledge source used, the value of a node for PageRank is proportional to its degree in undirected graphs. On the other hand, a significant difference between them is complexity. *Degree* is considered $O(n)$ and PageRank is $O(n^2)$, in other words, the time for the terms' analysis increases linearly and quadratically, respectively.

Another important finding regards external knowledge sources. Denser semantic relations, present in EnWordNet (an enhanced version of WordNet), increased the algorithms performance up to 9% (when in *lexical samples* modality). This increase is related to the fact that these approaches benefit from the number of relations to better

distinguish the importance of the vertices. The average number of relations exclusive to EnWordNet present in terms selected by the *Degree* algorithms is 20.5 edges. The original WordNet has hyperonymy and hyponymy as its most expressive relations, representing together the 9.29-edge average number in terms correctly selected by the *Degree* algorithm. Navigli and Lapata consider that, besides having a higher degree, the exclusive relations in EnWordNet establish important transversal connections and also those that are not necessarily part of taxonomy.

2.4. Graph Approach for WSD in Biomedicine texts

Studies that employ structured (graph) resources from external sources, in domain independent WSD, frequently use WordNet as structured knowledge source. Likewise, methods based on graphs were employed in specific domains, as it is the case of the Biomedicine domain.

Agirre et al. (Agirre et al. 2010) propose the use of the graph based approach for the Biomedicine domain. In this work, the personalized PageRank algorithm is employed in WSD, with the UMLS metathesaurus as knowledge source. The relations present in the UMLS are used for building a graph, which is then analyzed by the algorithm. Thus, the ranking of each candidate concept is generated based on its relative importance regarding the other concepts in the context of the ambiguous concept. This algorithm was previously used in a domain-independent context, employing WordNet as a knowledge base. It obtained better results than other proposals based on graphs, as analyzed by (Agirre and Soroa 2009).

Using the NLM-WSD corpus (Weeber et al. 2001), which is composed by abstracts on Biomedicine, the PageRank algorithm results were compared to the two baselines. Furthermore, the results were compared to those of (McInnes 2008), who used a subset of the NLM-WSD corpus. From these, around 54% are “difficult” cases according to Weeber et al. (2001). The relevance of the results obtained is one of the discussions pointed in (Agirre et al. 2010). Only 13 of the 50 concepts present in the NLM-WSD are related. This reduced set was initially established by Humphrey et al. (Humphrey and Rogers 2006), and was then used by McInnes (McInnes 2008). Humphrey et al. obtained most part of the best individual results. Around 76% of the concepts obtained the best result. The average achieved was 68.26% of correctness. McInnes’s approach did not obtain better individual results featuring a 48.11% average of correctness. Agirre et al.’s approach reached a 56.14% average of correctness. It obtained the best individual results in 3 concepts (around 23% of the total). In other two cases it was close to the best results. A highlighting fact is that Humphrey et al.’s approach uses a semi-supervised method. In comparison to the other unsupervised methods, the graph approach considerably increases the general performance of the WSD system in this reduced testing set.

3. Experiment

Considering the results from Agirre et al.’s experiment with an approach based on graphs in a specific domain, and Navigli and Lapata’s experiment on graphs in an independent domain, a new experiment is proposed. *Degree*, *KPP* and *PPR* were compared on Biomedicine domain.

In order to compare the results, the same requirements and means of interpretation used by (Agirre et al. 2010) were adopted. A set of instructions proposed by the authors, in addition to the material used, was collected from the website http://ixa2.si.ehu.es/ukb/ and other sources. All steps were performed taking into account the standard window of context (20 concepts, 10 before and 10 after the ambiguous one). The tools distributed by the authors run in two steps. In the first step the text file of the UMLS table should be used to generate a binary version of this table. The objective is to reduce execution time and optimize memory use. The second step uses binary version, a dictionary of concepts/CUIs, besides the ambiguous terms and their contexts. The results with the experiment's reproduction had a correctness percentage of 66.16%, considering 50 concepts. Similarly to how Agirre et al. (Agirre et al. 2010) positioned themselves, the annotated cases such as *none* are not part of this analysis.

The experiments performed with the two new algorithms, Degree and KPP, use the same framework, resources and parameters employed when reproducing the experiment with PPR.

3.1. Results

Agirre et al. (Agirre et al. 2010) presented a table with the disambiguation results for each concept in the corpus. Similarly to this, Table 1 contains the results excerpt of the two new algorithms proposed in our experiment. The concepts in italic represent the difficult cases according to (Weeber et al. 2001). The "#totalInst" column contains the amount of instances assessed by the algorithm for each concept. In other words, it represents all the instances of ambiguous concepts that were not classified as *none* by the annotators. The "#inst" columns show the amount of instances correctly classified by each algorithm. The percentage (%) columns present the rate of instances correctly classified, considering the "#inst" columns and the "#totalInst" column. In some cases more than one algorithm achieved the best result. The "Agirre et al. (2010)" column reproduces the percentage results that article presents for each concept. The absolute values of instances correctly classified are not shown.

Table 1. Result excerpt

Concept	#totalInst	Degree		KPP		PageRank		
		#inst	%	#inst	%	#inst	%	Agirre et al. (2010)
adjustment	93	13	13.98	15	16.13	29	31.18	35.50
cold	95	12	12.63	5	5.26	24	25.26	28.40
fit	18	0	0.00	18	100.00	2	11.11	11.10
reduction	11	2	18.18	2	18.18	5	45.45	54.50
resistance	3	3	100.00	3	100.00	3	100.0	66.70
secretion	100	97	97.00	1	1.00	99	99.00	99.00
transport	94	93	98.94	93	98.94	93	98.94	69.10
#inst sum	3983	1833		1676		2635		
Average			46.02		42.08		66.16	65.89

Among the 5000 corpus' instances, 3983 concept's instances were assessed. PPR correctly classified 2635 instances. From this total, only this algorithm reached the accurate sense of 580 instances (14.5%). KPP correctly classified 1676 instances and, among these, 676 (16.9%) were reached only with this algorithm. At last, 129 instances (3.2%) could be classified only by the Degree algorithm, which in its turn correctly classified 1833 instances. Some instances were correctly classified by more than one algorithm. In this situation, 458 instances (11.4%) were correctly classified by PPR and

KPP. A total of 1162 instances (29.1%) were correctly classified by PPR and Degree. KPP and Degree correctly classified 107 instances (2.6%) of the NLM-WSD corpus. Around 435 instances (10.9%) were correctly classified by all algorithms.

Ultimately, none of the three algorithms could correctly classify 436 instances (10.9%). Figure 2's diagram presents a distribution of the classification of instances in relation to algorithms.

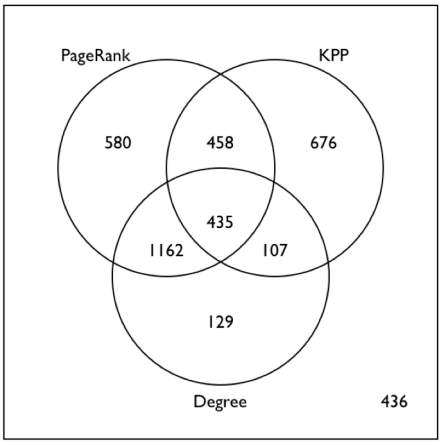

Figure 2. Distribution of the correctly classified instances per algorithm

3.2. Discussion

First, let's consider some particularities regarding the values listed on Table 1. The reproduction of Agirre et al.'s experiment led to a better result (66.16%) than the one reached with the experiment originally performed by the authors (65.89%). Two factors seem to account for this difference. The first one is that the parameters used by the UMLS table extraction tool might not be the same. There is no precise documentation about which vocabularies should be selected. The second one is that the UMLS may have received minor updates between the version used by the authors and the one used in the reproduction. Therefore, the graph structure and, consequently, the relations between concepts might be altered. The other algorithms did not achieve a better general result than the one reached by PageRank.

Another issue related to the results in Table 1 is the variation between the algorithms. Among the best results, in thirteen concepts (26% of the total) only one algorithm obtained a result higher or equal to double the other algorithms. For example, for the *fit* concept, KPP correctly classified 100% of the analyzed instances, and *PPR* only 11.1%. Contrary to what was presented in (Agirre et al. 2010), the determining factor for the classification choices is not the density with which the senses are connected. *KPP* distinguishes those that are central in the graph structure, besides the relation density. This behavior led to a contrary effect with *KPP*, where the *secretion* concept reached the worst result of the three algorithms (1% correctness). The *lead*, *resistance* and *transport* concepts had approximately 100% of the instances correctly classified by the three algorithms. In short, PPR obtained 62% (8) of the best results, while KPP reached 38% (5). *Degree* did not stand out in any of the concepts. The algorithms discussed in (Navigli and Lapata 2007; Navigli and Lapata 2010), which

outperformed on nonspecific domain, did not repeat such performance in the specific domain of Biomedicine.

All these aspects associated with result variations among algorithms and concepts led us to consider their performance at the instance level. If some algorithms can have very poor or very good results in relation to others, it is necessary to identify the proportion and the distribution of these results. Among those instances that were correctly classified, each algorithm's result for a corresponding instance allows to establish a set of thoughts. Firstly, *PPR* obtained the best general result (Table 1), *KPP* exclusively classified the highest number of instances. Those were 676 cases (16.94% of the 3983 instances) against 580 of the PageRank algorithm (14.56%). Notwithstanding, Degree correctly classified around 60% of the instances (1597) classified by PPR. However, according to (Navigli and Lapata 2010), the complexity of PPR and Degree is, respectively, $O(n^2)$ and $O(n)$.

4. Conclusions and future work

The experiment revealed that similar performances can be reached with different levels of complexity. Indeed, more than half of the instances can be analyzed in a shorter period of time if the Degree algorithm is used, for example. Furthermore, the possibility of identifying the most adequate algorithm to classifying a certain instance seems to be promising. A significant number of instances can be classified by just one of the algorithms, what identifies a correlation between instances and algorithms. On the other hand, a significant number of instances can be correctly classified by more than one algorithm. A study on this correlation between algorithms and instances is being developed as a continuation of this work. The objective is to identify instance features and a set of heuristics that allow the selection of the most adequate algorithm for the classification of words in the Biomedicine domain. Besides, we wish to identify and assess cases in which the right choice of algorithms with lower complexity can positively influence on performance.

References

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. Proceedings of the 12th Conference of the European Chapter of the ACL, 33–41.

Agirre, E., Soroa, A., & Stevenson, M. (2010). Graph-based Word Sense Disambiguation of biomedical documents. Bioinformatics, 26(22), 2889–2896.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. Computational and Mathematical Organization Theory, 12(1), 21–34.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), 107–117.

Garla, V. N., & Brandt, C. (2012). Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. Journal of the American Medical Informatics Association : JAMIA, 20(5), 882–6.

Haveliwala, T. H. (2002). Topic-sensitive PageRank. Proceedings of the Eleventh International Conference on WWW '02, 517.

Humphrey, S., & Rogers, W. (2006). Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. Journal of the American Medical Informatics Association, 57(1), 96–113.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: An Informatics Research Collaboration. Journal of the American Medical Informatics Association, 5(1), 1–11.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604–632.

McInnes, B. (2008). An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline. Proceedings of HLT-SRWS 2008, (June), 49–54.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of Biomedical Informatics, 46(6), 1116–1124.

Miller, G. a. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41.

Navigli, R. (2009). Word sense disambiguation. ACM Computing Surveys, 41(2), 1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. Lecture Notes in Computer Science (Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7147 LNCS, 115–129.

Navigli, R., & Lapata, M. (2007). Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), 1683–1688.

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(4), 678–92.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web.

Preiss, J., & Stevenson, M. (2013). DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples. In HLT-NAACL (pp. 1–4).

Trivedi, M., Sharma, S., & Deulkar, K. (2014). Approaches To Word Sense Disambiguation. International Journal of Engineering Research & Technology, 3(10), 645–647.

Weeber, M., Mork, J. G., & Aronson, a R. (2001). Developing a test collection for biomedical word sense disambiguation. Proceedings Annual Symposium. AMIA Symposium, 746–50.