Biomedical Text Disambiguation using UMLS

Wessam Gad El-Rab
University of Alberta
Edmonton, Canada
gadelrab@ualberta.ca

Osmar R. Zaïane
University of Alberta
Edmonton, Canada
Zaiane@ualberta.ca

Mohammad El-Hajj
MacEwan University
Edmonton, Canada
elhajjm@macewan.ca

Abstract—**Interest in extracting information from biomedical documents has increased significantly in recent years but has always been challenged by the ambiguity of natural language. An important source of ambiguity is the usage of polysemous words: words with multiple meanings. Word sense disambiguation algorithms attempt to solve this problem by finding the correct meaning of a polysemous word in a given context, but very few algorithms were designed to disambiguate biomedical text. In this study we propose a word sense disambiguation algorithm focused on biomedical text. The proposed algorithm does not need to be trained and uses a relatively small knowledge base.**

Keywords—Word Sense Disambiguation; UMLS; MetaMap

I. Introduction

Recent advances in biomedical research accompanied with the ease of electronic document creation have accelerated the rate of publishing electronic biomedical documents. Paradoxically, the resulting proliferation of biomedical documents increased the difficulty on health professionals being up-to-date with the latest medical findings. This led to the interest in automated tools such as information extraction (IE) and natural language processing (NLP) applied to the biomedical field.

Extracting information from biomedical documents is challenged by the ambiguity of natural language, in which words can have multiple meanings. For instance the word *"lens"* has different meanings in the following two article title sentences which we captured from the MSH-WSD dataset [14].

a) ***Lens*** cadmium, lead, and serum vitamins C, E, and beta carotene in cataractous smoking patients.

b) A simple solution to ***lens*** fogging during robotic and laparoscopic surgery.

In the first sentence, lens is used to refer to a *human body part*, while in the second sentence lens is used to refer to a *medical device part*.

Word sense disambiguation (WSD) is the process of finding the correct meaning "sense" of words that have multiple meanings. The correct sense of an ambiguous word can only be determined by analyzing the context in which the ambiguous word appears. For humans, the word sense disambiguation process is relatively easy as humans tend to do it unconsciously, while for machines it is as hard as an AI-complete problem, a technical term in artificial intelligence and complexity theory, which means solving it would require solving all the difficult problems in artificial intelligence (AI) such as natural language understanding [1].

There are many proposed approaches to address the WSD problem. For a classical comprehensive list of WSD algorithm classification, refer to [1] and for more recent studies refer to [2]. WSD algorithms at the highest level are classified either as *supervised learning* approaches or *unsupervised. Supervised learning* approaches must be first trained with a manually annotated corpus, while the *unsupervised* approaches do not require any annotated corpus and mostly rely on an external source of knowledge like a thesaurus or an ontology.

Generally, *supervised learning* approaches outperform *unsupervised* ones [3-6], but in the biomedical domain it is very expensive to create a manually annotated corpus for algorithm training purposes, which makes the unsupervised approach a more practical choice. In a WSD study focused in the biomedical domain [7] the authors believe that combining unsupervised learning and established knowledge proved to be most effective.

This paper presents an unsupervised graph-based approach to WSD in the biomedical domain that uses the unified medical language system (UMLS) [8] as its knowledge base. Section 2 describes the previous work on unsupervised graph-based WSD. Section 3 describes our unsupervised graph-based approach to WSD using the UMLS semantic network. Section 4 presents the evaluation of our algorithm. Finally, Section 5 concludes our findings.

II. Background and Related Work

Unsupervised graph-based WSD studies were mainly on generic domains, meaning that very few were specific to the biomedical domain; in Table I. we list six recent unsupervised graph-based WSD algorithms along with their domain, knowledge base, data sets used to evaluate the algorithm, and the reported accuracy. In the domain-independent WSD [10-13], we find that WordNet [15] is a commonly used

TABLE I. List of recent unsupervised graph-based WSD approaches

	Knowledge base	Evaluation Dataset	Accuracy
Bridget McInnes, Ted Pedersen, Ying Liu, Genevieve Melton (2011) [17]	UMLS Metathesaurus	MSH-WSD	72.0%
Eneko Agirre, Aitor Soroa, Mark Stevenson (2010) [9]	UMLS Metathesaurus	NLM-WSD	68.1%
Eneko Agirre, Aitor Soroa (2009) [10]	WordNet	Senseval-2 Senseval-3	58.6% 57.4%
Ravi Sinha, Rada Mihalcea (2007) [11]	WordNet	Senseval-2 Senseval-3	56.4% 52.4%
Roberto Navigli, Mirella Lapata (2007) [12]	WordNet EnWordNet	SemCor Senseval-3	--
George Tsatsaronis, Michalis Vazirgiannis, Ion Androutsopoulos (2007) [13]	WordNet	Senseval-2	49.2%

knowledge base, and Senseval [3-6], with its different versions, is the commonly used data set for algorithms evaluation. WordNet and Senseval can still be applied to biomedical text disambiguation but will result in lower accuracy when compared to a biomedical knowledge base and dataset. We can clearly see the difference when we compare the results between [10] and [9], where the authors applied the same algorithm, but used WordNet and Senseval in the first attempt [10] and UMLS and NLM-WSD in the second attempt [9] in which they achieved close to 10% accuracy improvement.

Since in our approach we use UMLS as our knowledge base and MetaMap as our concept mapping approach, we briefly present these two.

The UMLS is a repository of multiple controlled biomedical vocabularies developed by the U.S. National Library of Medicine (NLM) and is composed of the following three knowledge sources:

a) The *Metathesaurus,* a vocabulary database of biomedical concepts with their various names, and the relationships among them. The Metathesaurus of the UMLS 2011AB release contains more than 2.6 million concepts collected from 161 vocabularies.

b) The *Semantic Network,* a set of semantic types to categorize all concepts represented in the Metathesaurus, and a set of semantic relations to define possible relationships between semantic types. The Semantic Network in the UMLS 2011AB release contains 133 semantic types and 54 relationships.

c) The *SPECIALIST Lexicon*, a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech.

MetaMap is a program developed by the U.S. National Library of Medicine to map biomedical text to concepts in the UMLS. The algorithm of MetaMap parses the input text into phrases at the top-level. These phrases are decomposed into syntax units then into tokens at the lowest level. The algorithm does a lexical lookup of phrase words in the SPECIALIST lexicon then generates lexical variants of all phrase words. Subsequently, a matching process gets triggered to find matches between UMLS concept names and the generated lexical variants. The results are candidates and are ranked based on how well the UMLS concept matches the generated lexical variant.

III. Unsupervised Graph-based WSD

The algorithm we propose is based on the hypothesis that words closely located to each other in a text must have some degree of relatedness. We used the UMLS semantic network as our knowledge base to find the relatedness between words. In brief, for an ambiguous term *T* (i.e. for which we have different semantic types) we take the neighbouring words before and after in a given window and check their respective semantic types using MetaMap. We select the semantic type of *T* the one which has the smallest distance from the set of neighbouring word semantic types based on UMLS semantic network. Algorithm 1 shows the pseudocode of our approach.

ALGORITHM 1

Input:

1. *W,* a sequence of *n* words,
2. *t,* an index in *W* pointing to the word we need to disambiguate,
3. *s*, a window size of the words before and after *t* to include in the analysis.

```
WordSenseDisambiguate (W, t, s)
Load UMLS semantic network as a graph G
Map words $W_{1..n}$ to UMLS semantic types
let $A$ ={sematic types of $W_t$}
let $B$ ={sematic types of $W_l$ | $l$= ($t$-1..$t$-$s$) ∪ ($t$+1..$t$+$s$))}
for each a in A do
    for each b in B do
        RelatednessDist(a) ← RelatednessDist (a) + Shortestpath(a, b, G)
    end for
end for
let m = minimum{ RelatednessDist (a) | a in A}
return {a | a in A AND RelatednessDist (a) = m}
```

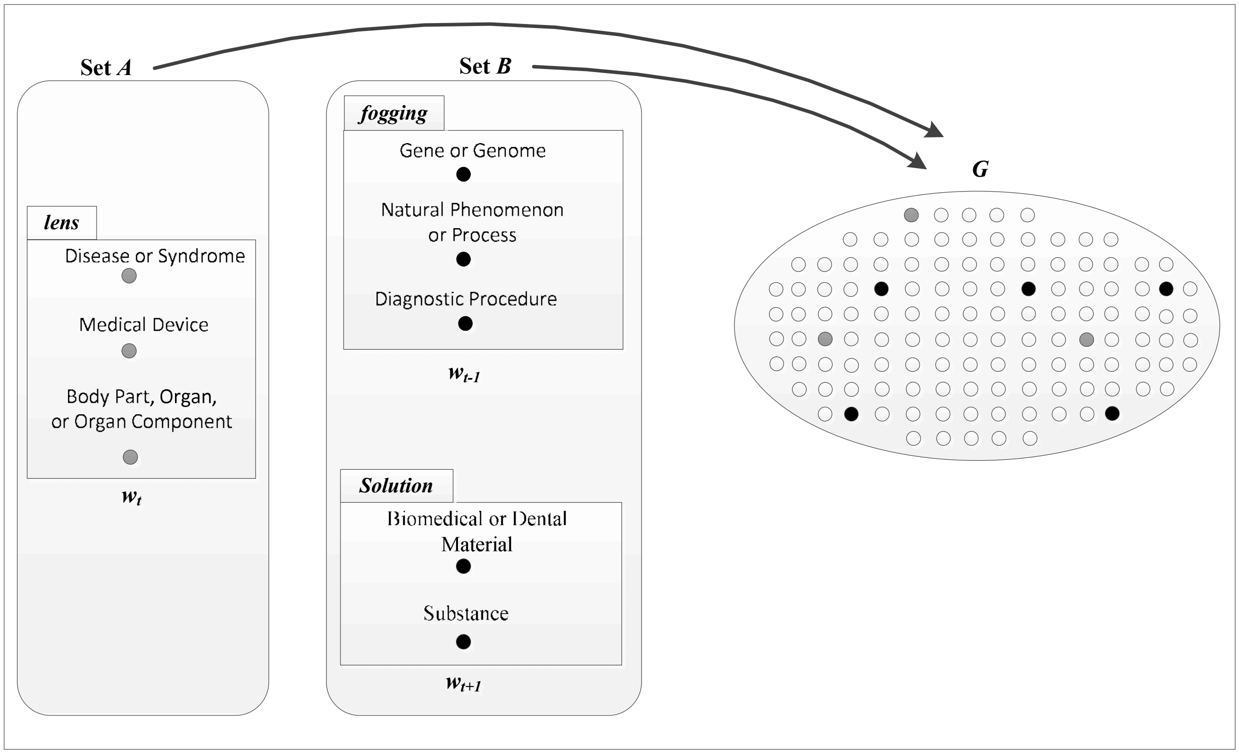

Fig. 1. Elements of set A, set B, and graph G

In line 1, we convert the UMLS semantic network to a directed graph G, where each semantic type is a node, and semantic relations between semantic types are the edges between the nodes. In line 2 we map all words in W to UMLS concepts using the MetaMap tool. In line 3 we populate set A with all the semantic types of the word we need to disambiguate W_t. In line 4 we populate set B with the semantic types of the words located before and after W_t by the given window size s. In lines 5-12 we measure the relatedness distance between each semantic type in set A to all semantic types in set B based on their closeness to each other in G. The semantic type of A that receives the lowest relatedness distance is deemed the semantic type of the correct sense. To prevent the algorithm to favour one central edge and consequently resulting in equal relatedness distances, we added weights to edges in G using the betweeness centrality [16].

As a running example we will use the following sentence from the MSH-WSD [14] data set.

> *A simple solution to lens fogging during robotic and laparoscopic surgery.*

The word we need to disambiguate is *lens*. As provided by the MSH-WSD data set the word *lens* can have any of the three possible UMLS concepts and their corresponding semantic types are shown in Table II. The correct concept of the word *lens* in the given sentence is C0023318 which has the sense of a medical device lens.

TABLE II. UMLS CONCEPTS

Concept		**Semantic Type**
Unique Id	**Name**	
C0023308	Lens Diseases.	Disease or Syndrome
C0023318	Lens (device).	Medical Device
C0023317	Lens, Crystalline.	Body Part, Organ, or Organ Component

Fig. 1 illustrates this running example. For simplicity, in the example we only take a size window of 1 and draw only 133 semantic types of graph G without the edges. We show elements of both set A and set B. Set A elements are the grey nodes representing the three candidate semantic types of W_t word *lens*, and set B elements are the black nodes representing the semantic types we extracted from MetaMap for the W_{t-1} word *solution* and the W_{t+1} word *fogging*.

We know that all the grey and black nodes of set A and set B must be nodes in the graph G, so we highlighted them in G. After having the graph G with highlighted grey and black nodes, the problem can be described as: which of the grey nodes is more related to the black nodes. To answer this question we calculate the sum of the shortest paths from each grey node to all the black nodes, and the grey node that receives the lowest values is deemed to have the highest relatedness.

IV. Evaluation

We evaluated our method using the MSH-WSD [14] dataset containing 203 ambiguous words. The 203 words are composed of 106 ambiguous terms, and 88 ambiguous acronyms, and 9 words that are combinations of both. The dataset has up to 100 instances for each possible sense. The total number of instances is 37,888. We ran our algorithm on the MSH-WSD dataset with a window of size 3 and the resulting average accuracy was 60.3%. Table III. shows the highest 10 accuracies and Table IV. shows the lowest 10 accuracies grouped by words. The small size window of 3 is chosen for scalability reasons as the semantic types of the neighbouring words add a combinatorial set of distances to compute. One important fact worth of note is that in our approach we use a relatively small knowledge base that we do not alter. We do not use the UML metathesaurus for instance. However, other methods use larger knowledge-bases encompassing larger UMLS knowledge sources, particularly the meta-thesaurus. Our approach is more extendable to take advantage of semantic relations.

TABLE III. Highest 10 accuracies

Word	True Positive	False Positive	False Negative	Accuracy
CDA	192	6	0	97%
CTX	177	6	0	97%
FAS	190	8	0	96%
MCC	124	7	0	95%
BPD	186	12	0	94%
BSE	186	12	0	94%
DAT	187	10	1	94%
Epi	187	11	0	94%
SS	136	7	1	94%
CRF	185	13	0	93%

TABLE IV. Lowest 10 accuracies

Word	True Positive	False Positive	False Negative	Accuracy
Lupus	12	0	285	4%
Medullary	8	0	190	4%
TPO	8	0	190	4%
TSF	2	0	51	4%
MBP	4	0	139	3%
TNC	5	0	162	3%
CCD	3	0	138	2%
RA	5	13	279	2%
Gamma-Interferon	1	0	197	1%
Murine sarcoma virus	0	0	180	0%

In our current analysis we used a graph to represent the UMLS semantic network, and set the edge weights independently of the semantic relation type. The UMLS semantic network has 54 different semantic relations; we believe that edge weights should be driven by the generalization level of the semantic relations to which it maps. As an example, a very specific semantic relation like the "*is a"* should be weighted higher than a general semantic relations like "*conceptually related*."

V. Conclusion

In this study we proposed a novel method that takes advantage of the UMLS semantic network to disambiguate terms in biomedical text. Our approach is more accurate than generic methods and is competitive with methods specifically designed for biomedical text. While the approach seems less accurate than previous specific methods, we cannot actually compare the methodology per se, as our approach relies on a smaller knowledge base and moreover is extendable to semantic analysis by taking advantage of semantic relations in the shortest path extracted from the UMLS semantic network.

Another avenue we plan to explore is investigating whether using dynamic windows size driven by the granularity of the semantic types of the word to disambiguate would improve accuracy.

References

[1] Ide, Nancy, and Jean Véronis.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational linguistics 24.1, 1998, pp. 2-40.

[2] Navigli, Roberto.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2, 2009

[3] Kilgarri, Adam.: Senseval: An exercise in evaluating word sense disambiguation programs. Proc. of the First International Conference on Language Resources and Evaluation,1998

[4] Edmonds, Philip, and Scott Cotton.: senseval-2: Overview. Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems. Toulouse, France, 2001, pp. 1-6

[5] Mihalcea, Rada, and Ehsanul Faruque.: Senselearner: Minimally supervised word sense disambiguation for all words in open text. In Proceedings of ACL/SIGLEX Senseval, vol. 3, 2004, pp. 155-158

[6] Agirre, Eneko, Oier Lopez de Lacalle, Bernardo Magnini, Arantxa Otegi, German Rigau, and Piek Vossen.: SemEval-2007 task 01: evaluating WSD on cross-language information retrieval. Advances in Multilingual and Multimodal Information Retrieval, 2008, pp. 908-917

[7] Schuemie, Martijn J., Jan A. Kors, and Barend Mons.: Word sense disambiguation in the biomedical domain: an overview. Journal of Computational Biology 12, no. 5, 2005, pp. 554-565.

[8] Humphreys, Betsy L., Donald AB Lindberg, Harold M. Schoolman, and G. Octo Barnett.: The Unified Medical Language System An Informatics Research Collaboration. Journal of the American Medical Informatics Association 5, no. 1, 1998, pp. 1-11

[9] Agirre, Eneko, Aitor Soroa, and Mark Stevenson.: Graph-based Word Sense Disambiguation of biomedical documents. Bioinformatics 26, no. 22, 2010, pp. 2889-2896

[10] Agirre, Eneko, and Aitor Soroa.: Personalizing pagerank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 33-41

[11] Sinha, Ravi, and Rada Mihalcea.: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity, In Proceedings of the IEEE International Conference on Semantic Computing (ICSC), 2007, pp. 363-369

[12] Navigli, Roberto, and Mirella Lapata.: Graph connectivity measures for unsupervised word sense disambiguation. In Proceedings of the 20th international joint conference on Artifical intelligence (2007) 1683-1688

[13] Tsatsaronis, George, Michalis Vazirgiannis, and Ion Androutsopoulos.: Word sense disambiguation with spreading activation networks generated from thesauri. In Proceedings of the 20th international joint conference on Artifical intelligence, 2007, pp. 1725-1730

[14] Antonio, Jimeno-Yepes, McInnes Bridget, and Aronson Alan.: Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics, 2011

[15] Stark, Michael M., and Richard F. Riesenfeld.: Wordnet: An electronic lexical database. In Proceedings of 11th Eurographics Workshop on Rendering, 1998

[16] Brandes, Ulrik.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25, no. 2, 2001, pp. 163-177

[17] McInnes, Bridget T., Ted Pedersen, Ying Liu, Genevieve B. Melton, and Serguei V. Pakhomov.: Knowledge-based Method for Determining the Meaning of Ambiguous Biomedical Terms Using Information Content Measures of Similarity." In AMIA Annual Symposium Proceedings, 2011, pp. 895