Unsupervised Graph-based Word Sense Disambiguation of Biomedical Documents

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*— Word Sense Disambiguation is the task of automatically identifying the correct sense of an ambiguous word. Biomedical documents, similar to other narrative documents, suffer from ambiguity, which impacts the ability to automatically extract knowledge contained in the document text. In this study, we propose a graph-based word sense disambiguation algorithm focused on biomedical text. The proposed algorithm uses the UMLS Metathesaurus as its source of knowledge. Evaluation is carried out using the MSH-WSD data set.**

Keywords—Word Sense Disambiguation; UMLS; MetaMap

I. INTRODUCTION

The increasing rate of published biomedical documents has negatively impacted health professionals' ability to keep up to date with the latest medical findings. This large amount of unstructured documentation has led to the interest in automated tools, such as information extraction (IE) and natural language processing (NLP) applied to the biomedical field.

Biomedical documents are written as narrative text, which is easy to interpret by humans, but very challenging for computers. One of the main challenges in extracting information from biomedical documents is the ambiguity of natural language, in which words can have multiple meanings. For instance the word *"astragalus"* has different meanings in the following two sentences, which we captured from the MSH-WSD dataset [14].

a) The biological course of fractures of the ***astragalus***.

b) Effects of ***astragalus*** injection on extracellular signal-regulated kinase pathway in cultured normal cardiac myocytes.

In the first sentence, *astragalus* is used to refer to a *human body part*, while in the second sentence *astragalus* is used to refer to a *plant*.

Word sense disambiguation (WSD) is the process of finding the correct meaning or "sense" of words that have multiple meanings. WSD resolves ambiguities by identifying the correct meaning of a word based on its context. Although research in this area has now been going on for decades, it is still considered a difficult problem. WSD 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 approaches to address the WSD problem for domain-independent text [1-2]. WSD algorithms can be distinguished as *supervised learning* or *unsupervised. Supervised learning* approaches use machine-learning techniques and require an annotated corpus for training, 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 impractical to manually annotate a corpus for algorithm-training purposes due to time and cost.

This paper presents an unsupervised graph-based approach to WSD in the biomedical domain that uses the Unified Medical Language System (UMLS) [7] as its knowledge base. Section 2 describes related work on WSD and also introduces the resources employed by the WSD systems used in this work. Section 3 describes our unsupervised graph-based approach to WSD using the UMLS Metathesaurus. Section 4 presents the evaluation of our algorithm. Finally, Section 5 concludes our findings.

II. BACKGROUND AND RELATED WORK

Most unsupervised WSD studies are domain ignorant, meaning that they are not customised for a specific field or domain. The key component that classifies an unsupervised WSD as domain specific is the knowledge base, for example the UMLS is commonly leveraged by WSD focused on the biomedical domain while WordNet [13] is commonly leveraged by domain-independent WSD. In Table I we list six recent unsupervised graph-based WSD algorithms along with their knowledge base, and the reported accuracy. As the reported accuracy shows, biomedical WSD achieve better accuracy compared to their domain-independent counterpart.

TABLE I. UNSUPERVISED GRAPH-BASED WSD APPROACHES

	Knowledge base	Accuracy
Bridget McInnes, Ted Pedersen, Ying Liu, Genevieve Melton (2011) [16]	UMLS Metathesaurus	72.0%
Eneko Agirre, Aitor Soroa, Mark Stevenson (2010) [8]	UMLS Metathesaurus	68.1%
Eneko Agirre, Aitor Soroa (2009) [9]	WordNet	58.6% 57.4%
Ravi Sinha, Rada Mihalcea (2007) [10]	WordNet	56.4% 52.4%
Roberto Navigli, Mirella Lapata (2007) [11]	WordNet EnWordNet	--
George Tsatsaronis, Michalis Vazirgiannis, Ion Androutsopoulos (2007) [12]	WordNet	49.2%

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) to support biomedical and clinical research. The UMLS 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, such as: SNOMED Clinical Terms (SNOMED-CT) and Medical Subject Headings (MSH). The Metathesaurus organises knowledge based on concepts, where each concept is identified by a Concept Unique Identifier (CUI).

b) The *Semantic Network,* a set of semantic types to categorise 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. Examples of semantic types include: Enzyme, Genetic Function, Therapeutic or Preventive Procedure, Laboratory Procedure.
- 54 semantic relations. Examples of semantic relations include: *affects*, *treats*, *disrupts*, *prevents*, *process_of*.

Semantic relations are interconnected by semantic types. For example, the semantic types *Enzyme* and *Genetic Function* are interconnected by the semantic relation *affects*.

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 and describes the morphologic, orthographic and syntactic properties of a word.

MetaMap is a program developed by the NLM to map biomedical text to concepts in the UMLS. The MetaMap system has five components:

a) Lexical/Syntactic Analysis: This component segments the biomedical text into phrases and then into terms. The text is Xerox part-of-speech tagged using the Xerox POS tagger.

b) Variant Generation: This component generates a variant for each phrase identified by the Lexical/Syntactic Analysis component. A variant is one or more phrase words accompanied with its spelling variants, derivational variants.

c) Candidate Identification: This component retrieves the set of concepts from the UMLS Metathesaurus that contain at least one variant identified by the Variant Generation component.

d) Candidate Evaluation: This component evaluates each candidate against the input text. The mapping score is computed using a combination of four linguistical measures: centrality; variation; coverage; and cohesiveness. The four measures are combined linearly such that coverage and cohesiveness get twice the weight of centrality and variation. The score is normalised to a value between 0 and 1,000, where a score of 1,000 means a perfect candidate.

e) Mapping Construction: This component combines all the Metathesaurus candidates that match the input text.

III. UNSUPERVISED GRAPH-BASED WSD

The algorithm we propose leverages the UMLS Metathesaurus as its knowledge base source. We represented the UMLS Metathesaurus as a graph *K*, such that UMLS concepts are the nodes and relation between UMLS concepts are the edges. The proposed algorithm is inspired by the approach presented in [11].

We used the MRCONSO table as the source of our nodes in the graph *K*, using the CUI as the node identity. The MRREL table is used as the source of the edges in the graph *K*. Table II shows a subset of relations between concepts that we extracted MRREL tables. The MRREL table contains ten different types of relations between concepts; for the performance consideration we focused on the following six relation types:

- PAR, the *parent* relation
- CHD, the *child* relation
- RB, the *broader* relation
- RN the *narrower* relations
- SIB, the *sibling* relation
- RO, the *other* relation

TABEL II . UMLS CONCEPTS REALATIONS

UMLS Concept	REL	UMLS Concept
Metabolisms, Energy	CHD	Rates, Basal Metabolic
Metabolisms, Energy	PAR	Processes, Metabolic
Drug-Induced Abnormality	RN	Fetal hydantoin syndrome
Drug-Induced Abnormality	PAR	Deformity
Drug-Induced Abnormality	RN	Warfarin syndrome

After building the knowledge source and represent as the graph K, it is fed into our algorithm along with the following inputs:

- W, a sequence of n words, representing the text containing the word to be disambiguated,
- t, an index in W pointing to the word we need to disambiguate,
- s, a window size of the words before and after t,
- A, a set of plausible senses for the word being disambiguated.

Algorithm 1 shows the pseudocode of our approach. We progressively build a graph for each W_t word to be disambiguated; the graph is composed of:

- V, a set of nodes representing the UMLS concepts of the words before and after W_t within a window of size s, combined with the set A. We used the MetaMap tool for mapping words to UMLS concepts. In line 3-9, we loop through all nodes in V, and for each node in V we search for its neighbour nodes in the graph K using depth-first search. All neighbour nodes found in K that do not exist in V are added to the V.
- E, the edges that interconnect all nodes in V based on the K graph.

ALGORITHM 1

WordSenseDisambiguate (*K, W, t, s, A*)
1: **let** V ={UMLS concept of $W_l \mid l = (t\text{-}1..t\text{-}s) \cup (t+1..t+s)$)}
2: **let** $V = V \cup A$
3: **for** each v in V **do**
4: X = DFS(K,v,p)
5: **for** each x in X **do**
6: **if** (x not in V)
7: **let** $V = V \cup \{x\}$
8 **end if**
9: **end for**
10: **end for**
11: **let** E = *GetEdges(V,K)*
12: **let** *VRanks = Betweenness(V,E)*
13: **let** m = *maximum*{ *VRanks* (a) | a in V and a in A}
14: **return** m

DFS(K,v,p)
1: return the set of nodes encountered when performing depth-first search starting from node v in the graph K at a maximum depth p.

GetEdges(V,K)
1: return the set of edges in graph K that interconnect all nodes in the V set.

Betweenness (V,E)
1: return a set of all nodes in V with their betweeness metric

We compute the betweenness score [15] of all nodes of the graph (V,E), the node in V that exist in A and receive the highest betweenness score is assumed to be the node of the correct sense of the W_t word.

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, 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 2 and the resulting average accuracy was 59.2%. Table III shows the highest 10 accuracies and Table IV shows the lowest 10 accuracies grouped by words.

TABLE III. HIGHEST 10 ACCURACIES

Word	True Positive	False Positive	False Negative	Accuracy
Lawsonia	99	16	0	86.09%
Eels	104	26	0	80.00%
HR	87	10	12	79.82%
DE	98	27	1	77.78%
PCB	93	28	6	73.23%
Torula	89	33	0	72.95%
PAF	82	33	0	71.30%
Callus	99	51	0	66.00%
EM	82	47	0	63.57%
CCD	88	42	11	62.41%

TABLE IV. LOWEST 10 ACCURACIES

Word	True Positive	False Positive	False Negative	Accuracy
Hemlock	19	54	4	24.68%
PCP	72	225	0	24.24%
CP	70	227	0	23.57%
Arteriovenous Anastomoses	30	99	0	23.26%
DON	26	100	0	20.63%
ORI	22	101	0	17.89%
MAF	21	99	0	17.50%
PCA	79	390	22	16.09%
WBS	17	111	0	13.28%
PHA	12	98	0	10.91%

V. CONCLUSION

In this study we proposed a word sense disambiguation algorithm that takes advantage of the UMLS Metathesaurus to

disambiguate terms in biomedical text. Our approach uses six relation types of the UMLS Metathesaurus, and builds a graph for each word to be disambiguated, where the graph's nodes get scored based on the betweenness metric, and out of the nodes that represent the different possible senses of the word being disambiguated, we assume that the node with the highest betweenness score is the node of the correct sense. The algorithm is evaluated using the MSH-WSD dataset and the resulting average accuracy was 59.2%. One avenue we plan to explore is analysing the impact of the different subsets of the UMLS Metathesaurus relations on the WSD algorithm 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] Agirre, Eneko, Aitor Soroa, and Mark Stevenson.: Graph-based Word Sense Disambiguation of biomedical documents. Bioinformatics 26, no. 22, 2010, pp. 2889-2896

[9] 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

[10] 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

[11] 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

[12] 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

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

[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] Freeman, L.C.: A Set of measures of centrality based on betweenness. Sociometry 40(1), 1977, pp. 35-41

[16] 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