Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2011) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare

Analyzing the impact of UMLS relations on word-sense disambiguation accuracy.

Wessam Gad El-Rab[a,*], Osmar R. Zaïane[a], Mohammad El-Hajj[b]

[a]*University of Alberta, Edmonton, Canada*
[b]*MacEwan University, Edmonton, Canada*

Abstract

Word-sense disambiguation (WSD) is the process of finding the correct meaning of words that have multiple meanings. The unsupervised WSD algorithm is the type of WSD algorithm that leverages an external source of knowledge to guide the disambiguation process. The unsupervised WSD algorithm type is attracting more interest in the biomedical domain because of its implementation practicality, especially when it leverages the knowledge sources of the Unified Medical Language System (UMLS), but still the resulted accuracy of the unsupervised WSD algorithm is lower than its supervised alternative. In this study we analyze the impact of using different subsets of the UMLS on the resulted accuracy of the unsupervised WSD algorithm. Our findings show that there are better ways to leverage the UMLS than using it as a monolithic source of knowledge.

Keywords: Word Sense Disambiguation, UMLS

1. Introduction

The proliferation of electronic narrative-based biomedical documents along with the need for accurate information retrieval systems have created a strong interest in automated tools such as information extraction (IE) and natural language processing (NLP) applied to the biomedical field.

* Corresponding author. Tel.: +1-780-700-8675.
E-mail address: gadelrab@ualberta.ca.

Information extraction tools are challenged by 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 data set [1].

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

b) Dietary supplementation with ***Astragalus*** polysaccharide enhances ileal digestibilities and serum concentrations of amino acids in early weaned piglets.

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

WSD is the process of finding the correct meaning of ambiguous words in context. The correct meaning – "sense" – of an ambiguous word can only be determined by analyzing the context in which the ambiguous word appears. WSD is categorized 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 [2].

There are many proposed approaches to address the WSD problem. For a classical comprehensive list of WSD algorithm classification, refer to [2] and for more recent studies refer to [3]. Fundamentally, WSD algorithms 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 such as a dictionary, thesaurus, semantic network, or ontology. The knowledge source leveraged by unsupervised WSD algorithms can be general-purpose, like the WordNet [4] thesaurus, or domain-specific, like the UMLS [5] biomedical thesaurus.

Our focus in this study is on the unsupervised WSD algorithms that leverage the UMLS as the knowledge source. There have only been a few attempts in this research area with different reported accuracies. Interestingly the difference in accuracy cannot only be credited to the rigorousness of the algorithm as each algorithm used different subsets of the UMLS, which could have a special impact. Moreover not all algorithms were evaluated using the same data set. The purpose of this study is to analyze the impact of the different subsets of the UMLS on the WSD algorithm accuracy. Section 2 provides the background information. Section 3 describes the previous work on unsupervised WSD using UMLS. Section 4 describes the WSD algorithm used for our analysis. Section 5 provides descriptions of the evaluation data sets. Section 6 discusses our analysis results. Finally, Section 7 concludes our findings.

2. Background

2.1. Unified Medical Language System

The UMLS [5] 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 repository of biomedical concepts, and the relationships among them. The Metathesaurus is considered the major component of the UMLS; the UMLS 2011AB release contains more than 2.6 million concepts collected from 161 vocabularies.

b) The *Semantic Network,* a set of categories – "semantic types" – used to categorize all concepts represented in the Metathesaurus. The Semantic Network also contains a set of relations –

"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.

2.2. MetaMap

MetaMap [6] is a program developed by the U.S. NLM to map biomedical text to the UMLS Metathesaurus. The algorithm of MetaMap is as follows: the input text is parsed into phrases at the top level, decomposed into syntax units, and then into tokens at the lowest level. For each phrase, a lexical lookup of words in the SPECIALIST lexicon is performed, then lexical variants of all phrase words are generated. 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.

3. Related work

There are multiple unsupervised WSD algorithms that leverage the UMLS. Some algorithms used the UMLS Metathesaurus knowledge source [7, 8], while others used the UMLS Semantic Network knowledge source [9, 10]. Generally, WSD algorithms that leverage the UMLS Semantic Network will run faster compared to the WSD algorithms that leverage the UMLS Metathesaurus, because of the smaller size of the Semantic Network knowledge base. But the main disadvantage of leveraging the UMLS Semantic Network is it restricts the WSD algorithm to only disambiguate words with concepts that belong to different UMLS semantic types. In the following subsections we provide a brief description of two different types of unsupervised WSD algorithms that used the UMLS Metathesaurus knowledge source.

3.1. Similarity-based unsupervised WSD

The similarity-based unsupervised WSD measures the similarity of each sense of the word being disambiguated to the words in the surrounding text, and the sense that has the highest similarity is assumed to be the correct one. The approach presented in [8] is a recent implementation of a similarity-based unsupervised WSD.

3.2. Graph-based unsupervised WSD

The graph-based unsupervised WSD builds a graph representing all possible senses of the word being disambiguated. The nodes in the graph correspond to the senses and the edges in the graph correspond to the relation type between senses (e.g. parent, child, broader). Next, the graph is assessed to determine the importance of each node: the node "sense" that is considered the most important of the word being disambiguated is assumed to be the correct one. The approach presented in [7] is a recent implementation of a graph-based unsupervised WSD.

4. Our approach

Our study focuses on analyzing WSD accuracy, and the way it is impacted by the different subsets of the UMLS Metathesaurus. For the purpose of our analysis we implemented a graph-based unsupervised WSD algorithm that computes the importance of each node *"sense"* in the graph using the PageRank

metric [11]. The algorithm is inspired by the approach presented in [12]; Algorithm 1 show the pseudocode of our approach. We ran the WSD algorithm against different subsets of the UMLS Metathesaurus. The way we split the UMLS Metathesaurus into smaller knowledge bases is by the different relations defined in the MRREL table, so each subset contains all the UMLS concepts but with only specific types of relations interconnecting them. We created four Metathesaurus subsets:

- PAR/CHD, a subset that contains only the *parent* and *child* relations;
- RB/RN, a subset that contains only the *broader* and the *narrower* relations;
- SIB, a subset that contains only the *sibling* relation;
- RO, a subset that contains only the *other* relation.

ALGORITHM 1

Input:

1. *K,* a graph representing the subset the UMLS Metathesaurus,
2. *W,* a sequence of *n* words,
3. *t,* an index in *W* pointing to the word we need to disambiguate,
4. *s*, a window size of the words before and after *t* to include in the analysis,
5. *A*, a set of plausible senses for the word being disambiguated. Only one element of *A* is the correct sense.

WordSenseDisambiguate (*K, W, t, s, A*)

1: **let** V ={UMLS concept of $W_l \mid l = (t-1..t-s) \cup (t+1..t+s)$)}
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 = PageRank(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.

PageRank (*V,E*)

1: return a set of all nodes in *V* with their PageRank metric

Each of the four UMLS Metathesaurus subsets is represented as a *K* graph*,* where the UMLS concepts are the nodes, and the UMLS relations between concepts are the edges. For the mapping step (line 1 of the WordSenseDisambiguate function), we used the MetaMap tool. In the DFS function we set *p* (the maximum depth of the depth-first search) to 1 for execution time purposes.

5. Evaluation data set

A majority of the WSD algorithms reported evaluations are based on one test data set; very few algorithms are evaluated on two or more test data set. The availability of different test data sets complicate the task of comparing the accuracy of the different WSD algorithms, as a WSD algorithm does not perform with the same reported accuracy on all other data sets, because each data set has different coverage of terms and concepts.

Selecting a data set for the purpose of evaluating a WSD algorithm is a critical task as it impacts our understanding of the strength and weakness of the WSD algorithm. Obviously the broader the coverage of data set the better, but as the test data set has to be finite, it becomes impossible to build a test data set that can cover all possible terms in all plausible contexts and therefore it is crucial to define a few key properties required in a data set to be considered as a proper test data set for the disambiguation task.

The main goal of a WSD algorithm is to properly disambiguate senses. Therefore richness of ambiguous term should be the most important property of the test data set – and not only that but to have equal distribution of the different senses of any ambiguous term. Below we provide a brief description of two data sets that are rich with ambiguous UMLS concepts:

- The NLM WSD [13] data set consists of 50 frequently occurring ambiguous terms from the 1998 MEDLINE baseline. Each ambiguous term in the data set contains 100 instances. The total number of instances is 5,000.
- The MSH WSD [1] data set contains 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 data set has up to 100 instances for each possible sense. The total number of instances is 37,888.

6. Results and discussion

We executed our algorithm against the MSH WSD test data set, with a window size of 2, and we executed the algorithm using the 4 subsets of the MRREL table (PAR/CHD, RB/RN, RO, SIB) as defined in section 4. For each run we captured the accuracy for all terms/acronyms of the MSH-WSD data set, of which we list the top 5 accuracies in Table 1-4.

Table 1. Highest 5 accuracies of the PAR/CHD relation

Term/Acronym	**PAR/CHD**	RB/RN	RO	SIB
dC	**94.44%**	51.01%	5.56%	50.51%
HCl	**93.94%**	49.49%	50.51%	66.16%
PCD	**93.94%**	49.49%	49.49%	36.36%
BPD	**93.43%**	0.00%	0.00%	50.51%
SCD	**92.93%**	0.51%	49.49%	50.00%

Table 2. Highest 5 accuracies of the RB/RN relation

Term/Acronym	PAR/CHD	**RB/RN**	RO	SIB
PHA	15.45%	**86.36%**	9.09%	15.45%
PAF	22.61%	**86.09%**	16.52%	**96.52%**
PCB	77.95%	**83.46%**	0.80%	68.50%
lymphogranulomatosis	19.33%	**82.35%**	**83.19%**	15.13%
DON	2.38%	**78.57%**	3.97%	76.19%

Table 3. Highest 5 accuracies of the RO relation

Term/Acronym	PAR/CHD	RB/RN	**RO**	SIB
HR	25.69%	0.00%	**88.07%**	22.94%
lymphogranulomatosis	19.33%	**82.35%**	**83.19%**	15.13%
sex factor	8.40%	0.00%	**71.76%**	29.01%
CDR	41.50%	33.33%	**68.03%**	40.82%
Callus	21.33%	2.00%	**66.00%**	36.00%

Table 4. Highest 5 accuracies of the SIB relation

Term/Acronym	PAR/CHD	RB/RN	RO	**SIB**
PAF	22.61%	**86.09%**	16.52%	**96.52%**
MCC	**86.26%**	3.05%	25.95%	**93.89%**
Eels	19.23%	4.62%	0.00%	**91.54%**
BAT	46.46%	50.00%	0.00%	**90.91%**
CAD	54.55%	49.49%	50.00%	**90.40%**

From our observation of the resulted accuracy of the 4 UMLS subsets, there is no real winner; each UMLS relation excels in disambiguating some terms/acronyms, and this indicates that using all relations of the UMLS MRREL table is not necessarily the best approach.

7. Conclusion

In this study we proposed a novel analysis that shows the impact of using different UMLS subsets as a knowledge source on the unsupervised type of WSD algorithms. We found that using the whole range of the UMLS relations defined in the MRREL table of the Metathesaurus is not necessarily the best approach. In fact a smaller subset of the relations will result in better accuracy. One avenue we plan to explore in the future is to try to identify automatically which subset of the UMLS Metathesaurus to use for the word being disambiguated.

References

[1] 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

[2] 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.

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

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

[5] 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

[6] Aronson A, Lang F: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010, 17(3):229

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

[8] 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

[9] S.M. Humphrey,W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. Journal of the American Society for Information Science and Technolology, 2006, 57(1):96–113.

[10] D. Alexopoulou, B. Andreopoulos, H. Dietze, A. Doms, F. Gandon, J. Hakenberg, K. Khelif, M. Schroeder, and T. Wachter. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics, 10(1):28, 2009.
[11] S. Brin and M. Page, "Anatomy of a large-scale hypertextual Web search engine," in Proc. of the 7th Conference on World Wide Web, Brisbane, Australia, 1998, pp. 107–117.
[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, pp. 1683-1688
[13] Weeber M, Mork J, Aronson A: Developing a test collection for biomedical word sense disambiguation. Proceedings of the AMIA Symposium, American Medical Informatics Association 2001, 746.