Developing a Test Collection for Biomedical Word Sense Disambiguation

Marc Weeber, PhD, James G. Mork, MSc, Alan R. Aronson, PhD

{weeber,mork,alan}@nlm.nih.gov

Lister Hill National Center for Biomedical Communications
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Ambiguity, the phenomenon that a word has more than one sense poses difficulties for many current Natural Language Processing (NLP) systems. Algorithms that assist in the resolution of these ambiguities, i.e. disambiguate a word, or more generally, a text string, will boost performance of these systems. To test such techniques in the biomedical language domain, there is the need for a test collection of disambiguated ambiguous strings. We report on the development of a Word Sense Disambiguation (WSD) test collection that comprizes 5,000 disambiguated instances for 50 ambiguous UMLS® Metathesaurus® strings.

INTRODUCTION

Consider the following sentences that include the word **cold** taken from three different MEDLINE® abstracts:[1]

(1) A greater proportion of mesophil micro-organisms were to be found during the **cold** months than in warmer months.

(2) In a controlled randomised trial we analysed whether the use of the term "smoker's lung" instead of chronic bronchitis when talking to patients with chronic obstructive lung disease (**COLD**) changed their smoking habits.

(3) The overall infection rate was 83% and of those infected, 88% felt that they had a **cold**.

The sense of the word cold is different in each sentence. Cold in sentence (1) is an indication of the temperature, in sentence (2) the acronym of chronic obstructive lung disease and in sentence (3) cold is a disease. The fact that a single word may have more than one sense is called *ambiguity*. In natural language, ambiguity occurs at many levels, e.g., lexical, structural, semantic, and pragmatic. Also, it pervades normal language use; humans have to disambiguate constantly (and subconsciously) in normal communication using textual and other types of context.

The general opinion is that language in more restricted environments, such as medical research, is more specific and straightforward; there is less ambiguity. This may well be the case, but ambiguity is still present as shown by the examples above. Additionally, the UMLS® Metathesaurus® [2], the largest medical thesaurus, has more than 7,400 ambiguous strings that map to more than one thesaurus concept [3]. The word **cold**, for instance, maps to six different UMLS concepts, three of which we used in sentences (1) – (3).

MEDICAL NLP AND AMBIGUITY

Medical NLP systems, generally designed to analyze medical texts for decision support or indexing purposes, have to deal with ambiguities in language. Columbia University's MedLEE system, originally designed for a small medical (and language) domain has been applied to different fields within medicine. One of the problems encountered when broadening the scope of such a system is the introduction of ambiguities. A term or word has different senses in different medical disciplines. MedLEE has some ad-hoc rules to deal with ambiguities, but there is a need for new, machine learning (ML) techniques and a good collection of training data [4].

The objective of the National Library of Medicine (NLM)'s Indexing Initiative is to investigate NLP methods whereby automated indexing techniques can partially or completely substitute for current (manual) indexing practices [5]. Error analysis of the indexing system shows that the major problems concern ambiguity of strings. Also, MetaMap, a text to concept mapping program [6, 7] is currently unable to disambiguate ambiguous concepts. The *DAD*-system, a concept-based tool for literature-based discovery in biomedicine [8,9] uses MetaMap for the processing of MEDLINE texts. In replicating Swanson's literature-based discovery of the involvement of magnesium deficiency in migraine [10], the *DAD*-system showed that the abbreviation **mg** might be interesting for treating migraine. However, the *DAD*-system is not able to distinguish between the UMLS concepts *Magnesium*

[1] The PubMed [1] ID's are 9477717, 9411973, and 9578931 respectively.

and *Milligram* for **mg**. This means that spurious information on milligram is included in the system's output [9]. In their recent study on UMLS concept indexing, Nadkarni *et al.* think a fully automatic procedure is not yet feasible, in part because of ambiguity problems [11].

Though there is clearly a need, the only research on biomedical word sense disambiguation are [12] and [4]. These two studies use rule-based approaches for a few cases in small domains. Recently, WSD has seen an upsurge of interest in computational linguistics, illustrated by a 1998 special issue of *Computational Linguistics*, Vol 24(1) and a 2000 special issue of *Computer and the Humanities*, Vol. 34(1/2). Additionally, there are the SENSEVAL workshops.[2] The time is ripe to test the newly developed algorithms in the biomedical language domain. Essential for testing the algorithms is a collection of manually disambiguated biomedical text strings for use as a gold standard. This paper reports on the development of such a WSD test collection.

EXTENT OF AMBIGUITY IN MEDLINE

To appreciate the amount of ambiguity present in MEDLINE, we processed the 409,337 citations added to the citation database in 1998. The processing consisted of finding UMLS concepts in the titles and abstracts of these citations by means of the MetaMap program. MetaMap chunks the sentences into (mostly noun) phrases that are mapped to UMLS concepts. In this experiment, we use the 1999 version of the UMLS. Table 1 displays some basic statistics.

Table 1: Mapping Results for 1998 MEDLINE.

No. of citations	409,337
No. of non-ambiguous phrases	30,514,468
No. of ambiguous phrases	4,051,445

We observe that 11.7% of the more than 34 million phrases result in more than one mapping to UMLS concepts, i.e. there is an ambiguous mapping. The differences between concepts are best depicted by the different semantic types that have been assigned to them. Studying the data, we observed three types of ambiguities: a) simple ambiguities in which a string maps to more than one UMLS concept (94.3% of all cases), b) lexical ambiguities (5.5%), and c) complex ambiguities (0.2%). See Table 2 for examples.

Table 2: Three Types of Ambiguities.

Type	UMLS concept	Semantic type
Simple: **activity**		
	Activity <1>	Finding
	Activity <2>	Daily or recr. activity
	% activity	Quantitative concept
Lexical: **reported**		
	Reporting	Health care activity
	Reports	Intellectual product
	Report <2>	Intellectual product
Complex: **reproductive health policies**		
	Reproduction	Organism function
	+ Health	Idea or concept
	+ Policies	Regulatory activity
	Reproductive Health	Occupation or discipline
	+ Policies	Regulatory activity
	Reproduction	Organism function
	+ Health Policies	Regulation or law

METHODS

Because complex ambiguities are both difficult and rare, and because lexical ambiguities should be resolved by better parsing strategies, we focus on simple ambiguities in the remainder of this paper. To disambiguate the strings we use human raters.

Selection of Strings

Based on the list of ambiguous UMLS strings, we have selected 50 highly frequent ones for inclusion into the test collection. They are tabulated in Table 3. Some highly frequent strings were not included because the concepts they are mapped to were either difficult to distinguish or the UMLS did not provide informative and consistent definitions and (hierarchical) relationships.

The second and seventh columns provide the strings' frequency of occurrence in the 1998 MEDLINE citations. Columns three and eight provide the number of different senses, or UMLS concepts to which a string maps. For some cases, we do not use all concepts available in the UMLS because we judged some of them to be too close in sense to make a practical distinction. Columns 4 and 9 tabulate the number of concepts we discarded for each string. For instance, MetaMap maps the string **depression** to three different UMLS concepts: *Depression motion, Depressive episode, unspecified*, and *Mental Depression*. The latter two concepts are very close in sense, so we decided to use only the second of the two, *Mental depression*,

[2] See http://www.sle.sharp.co.uk/senseval2/ for more information.

Table 3: Ambiguous Strings in the NLM's WSD Test Collection. The italicized ones are problematic to obtain a good agreement between raters. Excl R = rater excluded, and Excl S = number of senses excluded.

String	Occurrences	Senses	Excl S	Excl R
adjustment	2,596	4	2	
association	18,531	3		
blood pressure	6,713	4	1	
cold	2,448	6		
condition	24,891	3		
culture	20,635	3	1	
degree	17,419	3		*
depression	7,577	3	1	
determination	36,779	3		
discharge	5,072	3	1	*
energy	7,327	3	1	
evaluation	19,319	3	1	
extraction	10,831	3		*
failure	7,989	3		
fat	6,112	3		*
fit	3,591	3		
fluid	5,991	3		
frequency	16,244	3	1	
ganglion	580	3		
glucose	11,205	3		
growth	20,712	3		
immunosuppression	1,596	3		
implantation	4,170	3		*
inhibition	24,121	3		*
japanese	2,924	3		*
lead	9,880	3		
man	5,243	4		
mole	3,642	4	1	
mosaic	569	5		
nutrition	3,456	4	1	
pathology	4,373	3	1	
pressure	9,118	4	1	
radiation	5,822	3		
reduction	22,979	3		
repair	6,771	3	1	*
resistance	13,132	3		
scale	6,734	4		*
secretion	13,276	3	1	
sensitivity	16,173	4		
sex	7,214	4		*
single	29,311	3		
strains	15,873	3		
support	20,228	3		
surgery	22,539	3	1	*
transient	7,053	3		
transport	10,018	3		
ultrasound	5,704	3	1	
variation	10,431	3		
weight	12,857	3		
white	4,384	3	1	

since the UMLS vocabularies define this concept more clearly.

For each string, we have added the sense "none" which the raters can select when none of the available senses suit a particular instance. Following the depression example, there are two UMLS senses plus the "none" option which leads to an ambiguity of degree three (Table 3, columns 3 and 8).

The discussion on which strings to use for the test collection and which senses to include for each string took place in a team of 11, the authors plus eight other researchers at the NLM with various backgrounds in library sciences, linguistics, medical informatics, and medicine. The members of this group also served as raters who disambiguated the instances.

For every one of the 50 strings, we selected 100 instances at random from the 1998 MEDLINE collection. Almost all of these instances originate from different citations. Thus, there were 5,000 instances to be disambiguated.

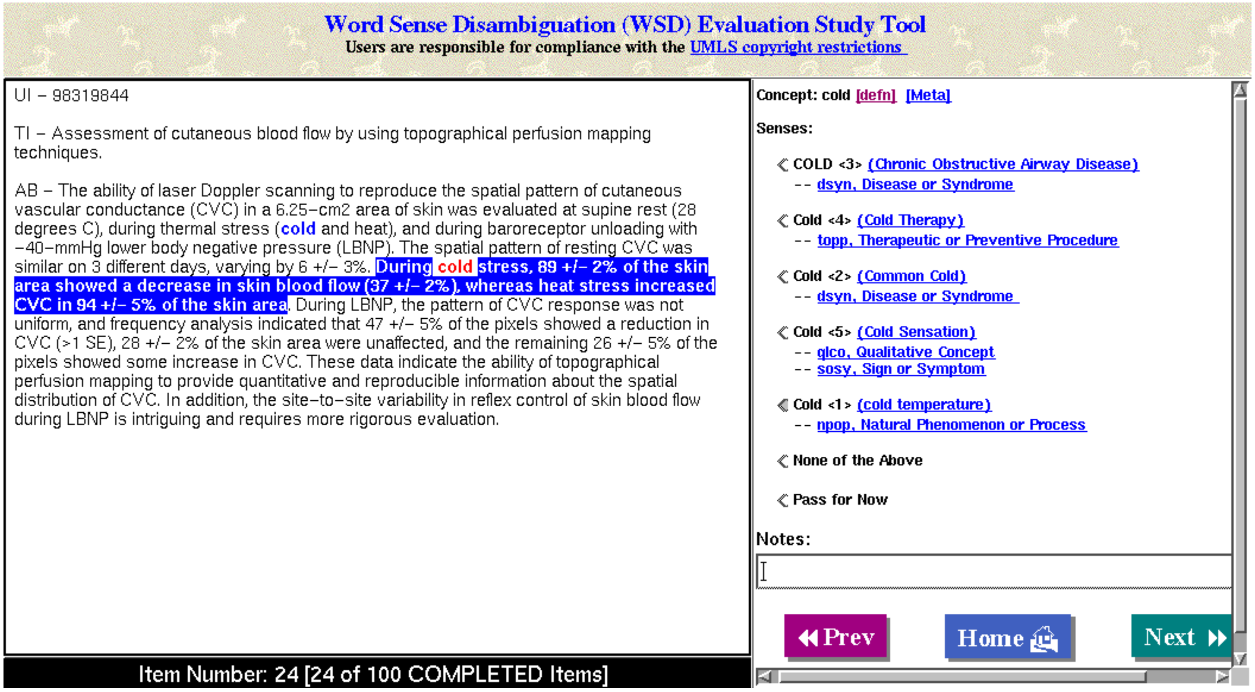
Disambiguation Procedure

Since disambiguating 5,000 instances of ambiguity manually is a non-trivial task, we developed a web-based interface that facilitates the disambiguation procedure and reduces the actual manual task to two mouse clicks for each instance, see Figure 1 for a screenshot.

The left panel of the interface presents the to be disambiguated string in red. The sentence in which it occurs, the direct context, appears in a blue box. Additionally, the rest of the title and abstract of the MEDLINE citation is visible. The raters were permitted to address the strings in any order and were not required to complete a string before starting another. The order in which the 100 instances for every string were presented had been randomized for every user. The different concepts (senses) are available in the right panel. The rater can only select one concept (radio button) or pass the instance to reconsider it at a later moment in time. Concepts and their semantic types have hyperlinks to the UMLS.

Analysis of Ratings

To reach a final classification on the correct sense, there are two approaches. The first one is majority voting. The sense that is selected by most raters will be the final and correct sense. The second method is latent class analysis (LCA) [13, 14]. This statistical method tries to find the underlying and "true" classifications.



Word Sense Disambiguation (WSD) Evaluation Study Tool

Users are responsible for compliance with the UMLS copyright restrictions

UI – 98319844

TI – Assessment of cutaneous blood flow by using topographical perfusion mapping techniques.

AB – The ability of laser Doppler scanning to reproduce the spatial pattern of cutaneous vascular conductance (CVC) in a 6.25–cm2 area of skin was evaluated at supine rest (28 degrees C), during thermal stress (**cold** and heat), and during baroreceptor unloading with −40–mmHg lower body negative pressure (LBNP). The spatial pattern of resting CVC was similar on 3 different days, varying by 6 +/– 3%. **During cold stress, 89 +/– 2% of the skin area showed a decrease in skin blood flow (37 +/– 2%), whereas heat stress increased CVC in 94 +/– 5% of the skin area**. During LBNP, the pattern of CVC response was not uniform, and frequency analysis indicated that 47 +/– 5% of the pixels showed a reduction in CVC (>1 SE), 28 +/– 2% of the skin area were unaffected, and the remaining 26 +/– 5% of the pixels showed some increase in CVC. These data indicate the ability of topographical perfusion mapping to provide quantitative and reproducible information about the spatial distribution of CVC. In addition, the site–to–site variability in reflex control of skin blood flow during LBNP is intriguing and requires more rigorous evaluation.

Item Number: 24 [24 of 100 COMPLETED Items]

Concept: cold [defn] [Meta]

Senses:

○ COLD <3> (Chronic Obstructive Airway Disease)
-- dsyn, Disease or Syndrome

○ Cold <4> (Cold Therapy)
-- topp, Therapeutic or Preventive Procedure

○ Cold <2> (Common Cold)
-- dsyn, Disease or Syndrome

○ Cold <5> (Cold Sensation)
-- qlco, Qualitative Concept
-- sosy, Sign or Symptom

● Cold <1> (cold temperature)
-- npop, Natural Phenomenon or Process

○ None of the Above

○ Pass for Now

Notes:

Figure 1: Disambiguation User Interface. The left panel shows the MEDLINE citation as context to the raters to disambiguate the string `cold`. The possible senses (concepts), with hyperlinks to the UMLS, are in the right panel.

This method may especially be useful when majority voting results in a tie. For any particular instance, LCA uses the rating patterns of the other instances to decide which is the true and final classification. In addition to these methods, it may be interesting to find out to what extent raters agree and disagree with each other using the kappa (κ) statistic [15].

The determination of the final classification is a four-step process. We repeated this process for all 50 strings. During step one, we compute the κ statistic for each rater–rater combination. This statistic shows which raters agree with each other, and more importantly, which raters disagree systematically from all others. We use the latter information in step two.

In step two, we count the total ratings for each instance of the string. If there is a majority of two votes for a certain sense, this will be the final classification. In case of ties, or many majorities of one, it may be interesting to exclude a rater if this rater disagrees systematically with all the others.

We apply step three if step two does not result in satisfactory results for many instances of the string, i.e. there are many ties and majorities of one and excluding one (or more) raters does not improve results. For these cases, we use LCA to obtain a classification.

Step four is the reassessment of instances in a group discussion of the disambiguation team. These instances did not obtain a reliable classification by step 2 or step 3.

RESULTS

Depending on the difficulty of the case, raters spent between thirty minutes and two hours per ambiguous string (100 instances). The rating task was done in addition to the raters' normal tasks. After a period of four months, during which there were three meetings in which the group discussed examples of difficult strings and particular instances, the data were frozen. Eight raters completed all the 5,000 instances, the other three completed 2800 (28 strings), 2200 (22 strings), and 600 (6 stings) respectively.

The agreement analysis by the κ statistic provided many interesting insights. For instance, the two raters who agreed best for most of the 50 strings are both former NLM indexers (the only two in the team). Also, for many strings, one or two of the raters disagreed systematically with the rest of the group. By excluding them in eleven cases (columns 5 and 10 in Table 3) we are able to resolve ties and many majorities of one. Eight raters were excluded at least once. Steps 1 and 2 were sufficient for 38 strings. Only 162 of the 3,800 instances had to be discussed in the team for a final classification (step 4). The twelve remaining strings, written in *italics* in Table 3, were more problematic in that there are many ties and majorities of one. After using LCA, still 159 of the 1,200 instances had to be discussed in the group to reach a final classification.

DISCUSSION

At the National Library of Medicine, we have developed a test collection for word sense disambiguation research. This collection will hopefully prove valuable for the future developments of medical NLP tools. As a first step we will apply different machine learning algorithms to disambiguate a string based on its context. The definition of the context will be one of the major challenges. The test collection provides the PubMed ID, the sentence in which the string occurs, the syntactic tags of the words in the sentence and the concepts that are found in the sentences by MetaMap [7]. Included with the concepts are their semantic types, therefore the semantic context may be included in the feature list that can be used by the algorithms.

We observe a distinction between two type of strings in the test collection: normal and problematic ones. For the problematic ones, it was difficult to obtain agreement among the raters on which sense is the accurate disambiguation of many of a string's instances. When human judgment is problematic, it may be impossible to automate disambiguation reliably. We therefore recommend to first consider the 38 normal strings (3,800 instances) with ML algorithms before turning to the problematic ones.

By Summer 2001, The WSD test collection will be available as a UMLS resource from the NLM at `http://umlsks.nlm.nih.gov/`.

ACKNOWLEDGMENTS

We would like to express our gratitude to the members of the disambiguation team: Florence Chang, Cliff Gay, Mike Hazard, Susanne Humphrey, Tom Rindflesch, Sonya Shooshan, Carolyn Tilley, and Sara Tybaert.

This research was supported in part by an appointment to the NLM Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the NLM.

REFERENCES

[1] National Library of Medicine. PubMed, 2001. `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed`.

[2] National Library of Medicine. Unified medical language system knowledge sources, 2001. `http://umlsks.nlm.nih.gov/`.

[3] Roth L, Hole WT. Managing name ambiguity in the UMLS metathesaurus. In: *Proc AMIA Annu Fall Symp 2000*. Philadelphia, PA: Hanley and Belfus, 2000; p. 1124.

[4] Friedman C. A broad-coverage natural language processing system. In: *Proc AMIA Annu Fall Symp 2000*. Philadelphia, PA: Hanley and Belfus, 2000; pp. 270–274.

[5] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM indexing initiative. In: *Proc AMIA Annu Fall Symp 2000*. Philadelphia, PA: Hanley and Belfus, 2000; pp. 17–21.

[6] Aronson AR. The effect of textual variation on concept based information retrieval. In: *Proc AMIA Annu Fall Symp 1996*. Philadelphia, PA: Hanley and Belfus, 1996; pp. 373–377.

[7] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The metamap program. In: *Proc AMIA Annu Fall Symp 2001*. Philadelphia, PA: Hanley and Belfus, 2001; *in press*.

[8] Weeber M, Klein H, Aronson AR, Mork JG, de Jong-van den Berg LTW, Vos R. Text-based discovery in biomedicine: The architecture of the *DAD*-system. In: *Proc AMIA Annu Fall Symp 2000*. Philadelphia, PA: Hanley and Belfus, 2000; pp. 903–907.

[9] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW. Using concepts in literature-based discovery: Simulating Swanson's raynaud – fish oil and migraine – magnesium discoveries. J Am Soc Inf Sci ;52 (7):548–557.

[10] Swanson DR. Migraine and magnesium: Eleven neglected connections. Perspect Biol Med 1988; 31 (4):526–557.

[11] Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases. J Am Med Inform Assoc 2001;8 (1):80–91.

[12] Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In: *Proc Annu Symp Comput Appl Med Care 1994*. Philadelphia, PA: Hanley and Belfus, 1994; pp. 240–244.

[13] Bruce RF, Wiebe JM. Recognizing subjectivity: A case study on manual tagging. Nat Lang Eng 1999;5 (2):187–205.

[14] Wiebe JM, Bruce RF, O'Hara TP. Development and use of a gold-standard data set for subjectivity classifications. In: *Proc ACL 1999*. Cambridge: The MIT Press, 1999; pp. 246–253.

[15] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76 (5):378–382.