

Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor
Indexing: Preliminary Experiment

Susanne M. Humphrey, Willie J. Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C.
Rindfleisch

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

E-mail: (humphrey, wrogers, halil, dina_demner, tcr)@nlm.nih.gov

ABSTRACT

An experiment was performed at the National Library of Medicine® (NLM®) in word sense disambiguation (WSD) using the Journal Descriptor Indexing (JDI) methodology. The motivation is the need to solve the ambiguity problem confronting NLM's MetaMap system, which maps free text to terms corresponding to concepts in NLM's Unified Medical Language System® (UMLS®) Metathesaurus®. If the text maps to more than one Metathesaurus concept at the same high confidence score, MetaMap has no way of knowing which concept is the correct mapping. We describe the JDI methodology, which is ultimately based on statistical associations between words in a training set of MEDLINE® citations and a small set of journal descriptors (assigned by humans to journals *per se*) assumed to be inherited by the citations. JDI is the basis for selecting the best meaning which is correlated to UMLS semantic types (STs) assigned to ambiguous concepts in the Metathesaurus. For example, the ambiguity *transport* has two meanings: “Biological Transport” assigned the ST Cell Function and “Patient transport” assigned the ST

Health Care Activity. A JDI-based methodology can analyze text containing *transport* and determine which ST receives a higher score for that text, which then returns the associated meaning, presumed to apply to the ambiguity itself. We then present an experiment in which a baseline disambiguation method known as MeSH Frequency was compared to four versions of JDI in disambiguating 45 ambiguous strings from NLM's WSD Test Collection. Overall average precision for the highest-scoring JDI version was 0.7873 compared to 0.2492 for MeSH Frequency, and average precision for individual ambiguities was > 0.90 for 23 of them (51%), > 0.85 for 24 (53%), and > 0.65 for 35 (79%). Based on these results, we hope to improve performance of JDI and test its use in applications.

INTRODUCTION AND BACKGROUND

Medical Text Indexer and MetaMap Application

The objective of NLM's Indexing Initiative (NLM Indexing Initiative, 2004) is to investigate methods whereby automatic indexing methods partially or completely substitute for current indexing practices (Aronson et al., 2000). The prototype indexing system developed under this initiative eventually became the Medical Text Indexer (MTI) (Aronson et al., 2004), which now actively participates in MEDLINE indexing using terms from NLM's MeSH® thesaurus (NLM Medical Subject Headings, 2004).. MTI indexes about 3,700 citations a day five nights a week. Indexers accept the option of viewing the resulting MTI recommendations about 379 times per day including weekends. It is estimated that MTI recommendations are accessed by indexers during

the indexing of 20% of MEDLINE articles. MTI has also been used as the sole indexing method for about 79,000 meeting abstracts on HIV/AIDS, health services research, and space life sciences.

MTI has as a major component the MetaMap program (Aronson, 2001), which maps biomedical text to concepts in the UMLS Metathesaurus (NLM Unified Medical Language System, 2004). MetaMap is a knowledge-based method that relies on the SPECIALIST Lexicon (a component of the UMLS) and an underspecified syntactic parser to identify noun phrases in biomedical text. The best match between a noun phrase and a Metathesaurus concept is computed by accommodating lexical variation in the input phrase and allowing partial matches between the phrase and concept. A confidence score is assigned to each mapping to reflect how closely the input noun phrase matches the target Metathesaurus concept. For example, the phrase *between the blastocyst trophoctoderm* in the following sentence from a MEDLINE abstract:

s1 In the mouse, the process of implantation is initiated by the attachment reaction between the blastocyst trophoctoderm and uterine luminal epithelium that occurs at 2200-2300 h on day 4 (day 1 = vaginal plug) of pregnancy.

maps to only one Metathesaurus concept:

694 Blastocyst [Embryonic Structure]

The confidence score, 694 out of 1000, and UMLS semantic type (ST) for the concept, Embryonic Structure, are provided as output. Semantic types are a set of 135 labels in the UMLS Semantic

Network for concept categories in the biomedical domain, e.g., Disease or Syndrome, Therapeutic or Preventive Procedure, Body Substance, and Pharmacologic Substance. Metathesaurus concepts are assigned one or more STs which form an “isa” link from the concept to the ST; in this example, Blastocyst isa Embryonic Structure.

However, the phrase *of implantation* maps to two Metathesaurus concepts, both with confidence scores of 1000:

1000 Implantation <1> (Blastocyst Implantation, natural) [Organism Function]

1000 Implantation <2> (Implantation procedure, natural) [Therapeutic or Preventive Procedure]

This result illustrates the problem of ambiguous mappings. Although “Blastocyst Implantation, natural” is the correct mapping, MetaMap has no way of choosing which of these concepts represents the meaning of this input phrase. This phenomenon is due to word sense ambiguity in English, and currently MetaMap does not choose between ambiguous mappings. Since MetaMap is the core component of MTI, automatic indexing of MEDLINE will be enhanced by providing a method for resolving this kind of ambiguity.

Word Sense Disambiguation Collection

The extent of the ambiguity problem was shown in an experiment conducted in connection with developing NLM’s Word Sense Disambiguation (WSD) test collection (Weeber et al., 2001) whereby 409,337 MEDLINE citations indexed in 1998 were run through MetaMap, resulting in

more than 34 million phrases. About 4 million phrases (11.7%) resulted in more than one mapping to Metathesaurus concepts; 94% of these cases were ambiguities in which an exact string maps to more than one concept. These sorts of ambiguity became the focus of developing the WSD test collection.

The purpose of the WSD test collection was to establish a testbed of humanly disambiguated instances to serve as a gold standard for evaluating automatic disambiguation methods. Based on the list of ambiguous strings from the processed phrases, 50 highly frequent ones were selected at random from the entire 1998 MEDLINE database. Appendix I shows all 50 ambiguities in the test collection with their respective Metathesaurus concepts and ST abbreviations. For example, the ambiguity *transport* maps to two concepts, “Biological Transport” with ST *celf* (abbreviation for Cell Function) and “Patient transport” with ST *hlca* (abbreviation for Health Care Activity). From now on will use abbreviated forms for the few STs mentioned in the text of this paper; their full forms can be looked up in Appendix II, which lists the 44 ST abbreviations and full forms represented in the test collection. Appendix III gives a hierarchical view of these STs.

For each ambiguity, 100 instances (sentences containing the ambiguity) were selected. Thus, there were 5,000 instances to be disambiguated by human raters. A Web-based interface was developed to facilitate the human disambiguation procedure, showing the citation with the highlighted sentence containing the ambiguous string to be considered. The actual manual task was reduced to two mouse clicks for each instance, these being the selection of one and only one sense or to pass for the time being. Figure 1 shows the result of the eight raters choices for disambiguating *s1*, unanimously in favor of “Blastocyst Implantation, natural” (having ST *orgf*).

PLACE FIGURE 1 HERE.

JDI-Based ST Indexing Applied to WSD

NLM is investigating Journal Descriptor Indexing (JDI), a novel approach to fully automatic indexing based on NLM's practice of maintaining a subject index to journal titles using terms, journal descriptors, corresponding to biomedical specialties (Humphrey, 1998; Humphrey, 1999). JDI methodology has been extended to ST indexing (Humphrey et al., 2000), both described in the next section. Using the above example, s1 can be indexed automatically by ST where each ST is ranked with a score from 0 – 1 (Figure 2). In this indexing, orgf (Organism Function) ranks higher than topp (Therapeutic or Preventive Procedure), thus indicating that “Blastocyst Implantation, natural” (having ST orgf) is a better meaning for the sentence than “Implantation procedure” (having ST topp), and therefore the better meaning for the ambiguous string *implantation* in this sentence, which is consistent with human raters (Figure 1).

PLACE FIGURE 2 HERE.

On the other hand, as seen in Figure 3, human raters unanimously selected “Implantation procedure” (having ST topp) for disambiguating the following sentence with the same ambiguous string *implantation*:

s2 We conclude that artificial sphincter implantation is safe, reliable and very effective in treating incontinence due to sphincteric dysfunction in properly selected patients.

ST indexing of s2 ranks topp higher than orgf (Figure 4), thus indicating “Implantation procedure” (having ST topp) is a better meaning for the sentence, and therefore the ambiguous string *implantation* in that sentence, also consistent with human raters (Figure 3).

PLACE FIGURE 3 HERE.

PLACE FIGURE 4 HERE.

This paper will describe experiments in applying JDI-based methodology to the WSD problem using the WSD Test Collection. This methodology will be explained in the next section.

METHODOLOGY OF JDI-BASED ST INDEXING

ST Indexing Using Word-ST Tables

Ultimately, JDI relies on ST indexing of some context in which the ambiguous string appears, as illustrated in the previous section where the context is the sentences containing *implantation*. If a sentence can be indexed by a ranked list of STs, and the ambiguous string in the sentence can be mapped to two possible concepts, each of which has a different ST assigned to it, then the higher ranked ST and its corresponding concept “win” as representing the meaning of the string. In other words, whichever ST ranks higher for the context of the ambiguity is considered the better of the two STs for the ambiguity itself; once the better ST is chosen, the corresponding concept is also chosen.

The ST indexing used for the WSD application relies on a word-ST table whereby each word in a training set is associated with an ST vector consisting of 129 ST rankings, ordered alphabetically by ST abbreviation. The training set consists of titles and abstracts of 910,542 MEDLINE citations to articles from 3,993 journals indexed in 1999 and 2000, which contain 232,676 unique words (meeting certain criteria such as having at least three characters, beginning with an alphabetic character, and occurring at least twice in the training set). Use of the JDI methodology for generating the word-ST tables based on the training set will be described further on. However, informally, an ST vector describes the semantic context in which a word occurs.

For example, ST vectors for the words *implantation*, *blastocyst*, and *sphincter* are shown in Figures 5, 6, and 7, respectively. Note: rather than display all STs, we selected the first and last STs (aapp [Amino Acid, Peptide or Protein] and vtbt [Vertebrate]) alphabetically by ST abbreviation, the set of highest ranking STs for each word (topp for *implantation*; emst [Embryonic Structure] for *blastocyst*; diap [Diagnostic Procedure] for *sphincter*), and the STs of interest for disambiguating *implantation* (orgf; topp) shown in boldface. High ranking STs in these examples reflect the semantic contexts in which the words commonly occur, and this has a significant impact on word sense disambiguation. *Blastocyst*, for example, most often occurs in text describing organism function, as seen by the high rank of the corresponding ST in Figure 6. *Sphincter*, on the other hand, is more often associated with procedures (high rank of topp in Figure 7). The two semantic types orgf and topp have relatively high rank in the ST vector *implantation* (Figure 5), which commonly occurs in both environments. As described subsequently, our methodology relies on computing semantic contexts for sentences containing ambiguous strings

like *implantation* by using pre-computed semantic contexts of co-occurring words in the sentence like *blastocyst* or *sphincter*

PLACE FIGURE 5 HERE.

PLACE FIGURE 6 HERE.

PLACE FIGURE 7 HERE.

Knowing the ST scores for individual words, we now can compute a vector which is the centroid of the ST vectors for all words in some context, such as a phrase or sentence. The score for an ST in the centroid is the average of the rankings for this ST across the words in the context. A display of STs in the centroid in rank order becomes the ranked ST indexing for the context. Figure 8 shows ST indexing for the phrase *blastocyst implantation* where the ST scores are the average of the same ST scores for *implantation* (Figure 5) and *blastocyst* (Figure 6). E.g., $(0.4998 [\textit{blastocyst} \textit{ orgf} \textit{ score}] + 0.6013 [\textit{implantation} \textit{ orgf} \textit{ score}]) / 2 = 0.5506 [\textit{blastocyst implantation} \textit{ orgf} \textit{ score}]$; *orgf* is appropriately ranked higher than *topp* for the phrase. Similarly, Figure 9 shows ST indexing for the phrase *sphincter implantation* where the ST scores are the average of the same ST scores for *implantation* (Figure 5) and *sphincter* (Figure 7); *topp* is appropriately ranked higher than *orgf* for the phrase.

PLACE FIGURE 8 HERE.

PLACE FIGURE 9 HERE.

The same methodology is applied for computing ST scores for the sentences containing the ambiguous string *implantation* in order to select the better concept mapping according to relative scores of STs assigned to the concepts. In ST indexing of s1 (Figure 2) the higher score for orgf (compared to topp) selects the “Blastocyst Implantation” concept, whereas in ST indexing of S2 (Figure 4) the higher score for topp selects the “Implantation procedure” concept.

JDI Methodology for Generating Word-ST Tables

JDI Indexing of Words

We will now describe the JDI methodology and how it is used for generating word-ST tables used for ST indexing. JDI uses statistical associations between the words in the training set and 127 journal descriptors (JDs) which index the approximately 4000 MEDLINE journals *per se* in terms of biomedical disciplines (NLM, 2002). Figure 10 shows a sample journal record (Journal Identifier, Title, Title Abbreviation, Journal Descriptor) for *Fertility and Sterility* in NLM’s journal (i.e., serial records) database.

PLACE FIGURE 10 HERE.

Figure 11 shows a sample citation (PubMed Identifier, Title, Title Abbreviation, Journal Identifier, Source, Journal Descriptor) from the training set, including the JD Reproduction, which we mapped from the journal record. Thus, citations inherit JDs from journal records corresponding to the journals in which the documents are published. Each word in the sample title

(Figure 11) from the training set (including *implantation*, which we emphasize) can be said to co-occur with the JD Reproduction by virtue of this inheritance.

PLACE FIGURE 11 HERE.

Since each citation in the training set inherits one or more JDs, an association between words and JDs can be represented as the number of co-occurrences of each word with each JD in the citations in the training set. The JD scores for *implantation* can be expressed by the ratio of the number of citations in which *implantation* co-occurs with the JD, divided by the total citation count for *implantation*. The 127 JD scores for *implantation*, ordered alphabetically by JD, form a JD vector. For example, part of the JD vector for *implantation* is shown in Figure 12. Note: rather than display all JDs, we selected the first and last JDs alphabetically (which, incidentally, never co-occur with *implantation*) and the five highest ranking JDs.

PLACE FIGURE 12 HERE.

We therefore can assign JDs as indexing terms to some text based on the words in it. Analogous to ST indexing which uses ST vectors, we perform JD indexing by computing a JD vector, which is the centroid of the JD vectors for the words in the text to be indexed. The score for a JD in the centroid is the average of the scores for this JD across the words. A display of JDs in the centroid in rank order becomes the ranked JD indexing for the text. Figures 13 and 14 show the first five JDs in the indexing of s1 and s2, respectively. The JD scores for each JD are the average of the scores for the same JD for words in the sentences. For example, for s1, the score for Reproduction is based on the average of the scores for Reproduction in the JD indexing of words taken from the sentence: *implantation, attachment, blastocyst, uterine, luminal, epithelium,*

vaginal, plug, pregnancy (allowing for conditions to ignore certain words, such as membership in a stopwords list and non-occurrence in the UMLS Metathesaurus). As shown in Figure 13, the outstanding JD for s1 is Reproduction; in Figure 14, the outstanding JD for s2 is Urology.

PLACE FIGURE 13 HERE.

PLACE FIGURE 14 HERE.

Creation and JD Indexing of ST Documents

However, this JD indexing as such isn't useful for WSD. What we need is ST indexing for selecting the best MetaMap concept mapping, as described earlier. The way we achieve this is by creating "ST documents" as documents to undergo JD indexing, where an ST document is a set of Metathesaurus words highly associated with a particular ST. An ST document is created by automatically extracting one-word Metathesaurus strings belonging to concepts assigned the ST; this set of words comprises the ST document. For example, the 2002 Metathesaurus contained 187 words in our "orgf document" (autoregulation, deglutition, healing, locomotion, urination, etc., where these words belonged to concepts assigned the ST Organism Function) and 1478 words in our "topp document" (arthroplasty, bandaging, dissection, hemodialysis, immunization, etc., where these words belonged to concepts assigned the ST Therapeutic or Preventive Procedure). Part of the JD vector for the latter ST document is shown in Figure 15, consisting of the five highest ranking JDs and the first and last JD alphabetically. We performed JD indexing of 129 ST documents (remaining STs did not have enough Metathesaurus words associated with them), resulting in a JD vector for each of them.

PLACE FIGURE 15 HERE.

Similarity between Word JD Vectors and ST Document JD Vectors

Using the standard vector cosine coefficient (Salton & McGill, 1983), we then computed the similarity, on a scale of 0 – 1, between the JD vector for each word in the training set and the JD vector for each ST document. Each word and its scores indicating similarity to ST documents (in terms of JD indexing), ordered alphabetically by ST abbreviation, became an entry in the word-ST table (i.e., an ST vector) used for ST indexing, as described earlier.

Looking again at Figures 5, 6, and 7, we now can interpret the items in these ST vectors in terms of similarity to ST documents. That is, JD indexing of *implantation* is more similar to JD indexing of the topp document than the orgf document; JD indexing of *blastocyst* is more similar to JD indexing of the orgf document than the topp document; JD indexing of *sphincter* is more similar to JD indexing of the topp document than the orgf document. Thus, ST indexing selects topp when the ambiguous string *implantation* occurs in a context (e.g., s1) containing words with JD indexing more similar to the topp document; conversely, ST indexing selects orgf when *implantation* occurs in a context (e.g., s2) containing words with JD indexing more similar to the orgf document.

RELATED WORK

Word sense disambiguation is a difficult but crucial task in many areas of automatic language processing, such as information retrieval (Clough & Stevenson, 2004; Vorhees, 1998), machine translation (Brown et al., 1991), and question answering (Pasca & Harabagiu, 2001). Beginning in the late 1950's, numerous solutions to the ambiguity problem have been explored. The growing interest in disambiguation methods and their performance led to formation of SENSEVAL, an international organization devoted to evaluation of word sense disambiguation systems. (Kilgarriff & Rosenzweig, 2000; Edmonds & Kilgarriff (2002); Mihalcea et al., 2004). For a review of existing disambiguation methods, which is beyond the scope of this paper, see Ide & Véronis (1998). Below we present work related to JDI either because of the similarity in the approach, or the common domain and collection used in the experiments.

The JDI method described in this paper combines a statistical, corpus-based method (two year MEDLINE training set) with utilization of pre-existing medical domain knowledge sources, JDs (NLM, 2002) and STs (NLM Unified Medical Language System, 2004).

Statistical methods are based on the idea that the given context determines the sense of the word. These methods rely on learning disambiguation rules from large sense-tagged corpora. Further distinction in the learning methods is based on the manner in which the text collection is annotated with word senses. Supervised methods that show the best performance in many natural language processing tasks rely on extensive high quality manual sense tagging of large amounts of text. This dependence restricts application of supervised methods to tasks and domains for which resources exist. Bootstrapping the annotation process with a smaller amount of hand tagged data, or resorting to fully automatic unsupervised methods has been suggested as a way to overcome the

data acquisition problem. (Yarowsky, 1995) Approaches that attempt to obtain annotated data, but avoid manual annotation have been explored recently. These methods include creating a collection by formulating a query using WordNet definitions of word senses, and searching the Web (Mihalcea & Moldovan, 1999); elicitation of volunteer contributions using a Web-based application (Mihalcea et al., 2004); and using text in parallel translations (Resnik, 2004),

In the spirit of avoiding costly manual annotation the JDI method assigns JDs and subsequently STs to the text in the training set thus avoiding the need to discover word senses in untagged text as in clustering-based unsupervised approaches (Schütze, 1992; Pedersen & Bruce, 1997; Pantel & Lin 2002). Since JD assignment and the subsequent steps are performed automatically, JDI is a rather sophisticated unsupervised approach that creates a representation of word senses (word-ST vectors) using co-occurrences of words with JDs (word-JD vectors) from the training set with the help of ST assignments to concepts in the UMLS Metathesaurus. Thus, the WSD collection is not used for training.

Using the UMLS and JDs as the source of knowledge is conceptually close to domain-independent methods that use pre-existing knowledge repositories, such as machine-readable dictionaries or thesauri for the same purpose. Dictionary-based methods, pioneered by Lesk (1986), compare the dictionary definitions of the word senses with the words in the context. These methods differ in types of sources used and the ways in which similarity between the sense representation and the word context is measured, and in general don't have the benefit of the sense assigned to the training set provided by JDs. Yarowsky (1992) developed a statistical model based on categories of Roget's International Thesaurus and text of the Grolier encyclopedia. Liddy &

Paik (1993) and Liddy et al. (1993) use Subject Field Codes (SFCs) from Longman's Dictionary of Contemporary English (LDOCE); however, the codes are manually assigned to each word in the dictionary by lexicographers rather than being propagated as in the JDI approach.

Domain Driven Disambiguation (Magnini et al., 2002) augments WordNet (Fellbaum, 1998) with domain labels from the Dewey Decimal Classification to represent the context and the word senses using domain vectors. Interestingly the kernel based system that incorporates this method was one of the best performing systems in the SENSEVAL-3 English lexical sample WSD task (Strapparava et al., 2004). This task, which requires annotation of instances of sample words in short extracts of text, is equivalent to the goal of the JDI method in disambiguating MetaMap output. It may be of interest to note that the average precision of JDI ranging from 77.10-78.73% depending on context (Table 1, in results and analysis section) is comparable to the precision of the top-performing supervised system participating in this SENSEVAL-3 task which is 79.3% (Mihalcea et al., 2004).

Maynard & Ananiadou (2000) use the UMLS and Semantic Network and the strength of association between a multi-word term and its context to identify one sense for that term in the corpus. Here again the JDI indexing of the training set permits finer granularity of the sense assignment, i.e. the word can be disambiguated given a paragraph, or a sentence.

The idea to disambiguate terms in the biomedical context using the UMLS semantic types of unambiguous neighboring concepts was introduced by Aronson et al. (1994). The availability of an extensive knowledge source such as UMLS has potential to significantly reduce or even

eliminate the need for manual sense annotation. One such unsupervised approach was studied by Widdows et al. (2003) who augmented information about concepts and semantic types with information about co-occurring concepts also contained in UMLS. In this approach, first all possible senses are found for each ambiguous word. Then all conceptually related and co-indexing terms for each sense are extracted from the corresponding sources (conceptually related terms can be found in the UMLS MRREL and MRCXT files, and the UMLS MRCOC file contains the co-indexing terms). Then the local context of the ambiguous word is examined for the presence of the related concepts. The sense that is supported by the largest number of related terms in the context is assigned to the ambiguous word. This study found both precision and recall to be better when only co-indexing terms were used for disambiguation as opposed to the combination of the co-indexing and hierarchically related terms. In another unsupervised approach Liu et al. (2002b) used the MRREL file to automatically annotate related concepts in MEDLINE citations. The presence of conceptual relatives permitted determining the sense of the ambiguous word in a large number of citations. The remaining citations were disambiguated using a naïve Bayes classifier trained on the previously disambiguated texts.

Since both unsupervised methods described above rely on the presence of related concepts in the citation, they might be sensitive to the exact wording of the text in the same manner that the early methods that used machine-readable dictionaries as the knowledge source were sensitive to the wording of the sense definitions. The advantage of the JDI method is that it does not require having specific words in the text containing the ambiguity (i.e., all words are pre-labeled with JDs inherited by the training set documents from the journals they appear in, and then labeled with STs

according to the methodology explained in the previous section), and thus it is not necessary to have large numbers of examples with these specific words.

Although our method is not supervised, it is important to mention two experiments that used parts of the NLM's WSD collection for supervised word sense disambiguation. Liu et al. (2004) studied various sizes of immediate contexts to the right and to the left of the ambiguous word for training of machine learning algorithms that demonstrated high accuracy in general English word sense disambiguation, namely naïve Bayes, decision list, and a combination of a naïve Bayes and an instance-based classifier. Since none of the classifiers in this experiment outperformed the rest for all ambiguities, the authors recommend selecting the best classifier individually for each term, and using supervised WSD only when there are at least a few dozen instances tagged for each sense of the word. Leroy & Rindfleisch (2004) studied the possibility of reducing the size of the required training set by utilizing symbolic knowledge encoded in the UMLS. In this experiment a naïve Bayes classifier was trained on sentences containing ambiguous words that were represented using a combination of syntactic features, semantic types found in the sentence, and semantic network relations, such as part-of, between these semantic types. We compare the performance of JD to these methods in the results and analysis section.

EXPERIMENTAL METHOD

Word Sense Disambiguator Tool

A Word Sense Disambiguator interface has been developed to determine the performance of individual disambiguation methods on the WSD Test Collection (Figure 16). This interface was used for running the baseline MeSH Frequency method (described below) and the JDI method to be compared to it. We have used Disambiguator in an experiment to measure the performance of MeSH Frequency and four versions of JDI corresponding to different contexts in which the ambiguity occurs, as described later in this section.

PLACE FIGURE 16 HERE.

MeSH Frequency Baseline

MeSH Frequency uses frequency counts of MeSH indexing term in a subset of MEDLINE citations. (MeSH Frequency forms the baseline for developing JDI, but is not used in an implemented system.) Each candidate concept for an ambiguity is matched to a MeSH synonym, if there is one. The concept that has the MeSH synonym with the highest frequency count in MEDLINE is returned as the disambiguator answer. Figure 17 shows the first few lines of the results for MeSH Frequency in disambiguating the instances of the *implantation* ambiguity discussed in previous sections of this paper. (Only 67 instances are processed as a training set for disambiguation methods; the remaining 33 are reserved as a test set.) In a line of results, the Item ID identifies the ambiguous text. For example, in the last line of Figure 17, 9344537.ab.1 stands for the first sentence in the abstract in the citation with PMID 9344537. Next on the line is the reviewed answer from the consensus of human raters, followed by the disambiguator answer for the particular method that was selected, in this case Word Frequency. Clicking on this Item ID

displays the citation with the sentence containing the ambiguity highlighted (Figure 18). This display is similar to the one shown to human raters in developing the WSD Test Collection. Also highlighted is the ambiguity in other sentences, although raters focused on the highlighted sentence for the disambiguation. This display is informative in evaluation of automatic indexing methodologies by allowing viewing of the context of the ambiguity. The ambiguous text in Figure 18 is our sample s1 sentence.

PLACE FIGURE 17 HERE.

PLACE FIGURE 18 HERE.

Referring to Figure 17, for *implantation*, the MeSH Frequency method selects “Blastocyst Implantation, natural” as the correct concept for all 67 instances. This is the reviewed answer for only 11 instances, and is reflected in the TP (True Positive) number in the Overall Summary line. Precision in this line is the precision score of 0.1642, which is TP / Count (total count of 67). The reason for this poor performance is that this concept has a MeSH synonym (Ovum Implantation) but the other concept “Implantation procedure” has no MeSH synonym. The Overall Summary also gives counts and scores ignoring the instances where “None of the Above” is the reviewed answer. For this ambiguity, there was only one “None of the Above”; therefore, ignoring this instance, $\text{Count} = 66$, and $\text{Precision} = 11/66 = 0.1667$. We are using scores ignoring “None of the Above” because neither MeSH Frequency nor the JDI method is designed to return this answer (see discussion about this at the end of this section).

As shown in Table 1, the average score for MeSH Frequency is 0.2491, which is the average of the precision scores for the 45 ambiguities processed by this method in the experiment (see

discussion on elimination of five ambiguities at the end of this section). Practically half the ambiguities have a precision score of 0.0000 (the disambiguator answer is “No match found” for all instances) on account of absence of MeSH synonyms for all candidate concepts. In cases where performance is good for this method, the concept having the MeSH synonym with the highest frequency happens to be correct for most instances.

Contexts Evaluated in Experiments

A particular methodologic issue that arises for the JDI method is what should be the context for an ambiguous instance. Should it be just the sentence in which the ambiguous string appears (i.e., target sentence)? Should it be the entire citation? An alternative context for the citation is the target sentence together with other sentences containing the ambiguity, or morphological variant of the ambiguity. Variants were determined using the UMLS SPECIALIST Lexicon; for example, variants of the ambiguous string *culture* are *cultures*, *cultured*, *culturing*, *cultural*. A question arose in the situation where the desired context is all sentences with the ambiguity/variants, but there is only one sentence that qualifies, i.e., the one with the ambiguity. Is some additional context always desirable beyond this sentence? We therefore derived a rule that if this sentence has fewer unique words than some threshold, the system would go to the entire citation as context. Figure 19 summarizes the contexts in our preliminary experiments.

PLACE FIGURE 19 HERE.

Results of JDI using the various contexts for the 45 remaining ambiguities will be presented in the results and analysis section for comparison with one another and with MeSH Frequency.

Problematic Issues

Five of the ambiguities were eliminated for this experiment: *association*, *cold*, *man*, *sex*, and *weight*. The last four of these are each mapped to two concepts having the same ST. For example, *weight* is mapped to the concepts Body Weight and Weight, both of which are assigned the ST qnco (in addition, Body Weight is mapped to orga); for the more than 40 instances where JDI found qnco to be the better ST (over orga), the system had no way of knowing which of the two concepts to select, since they were both assigned this same ST.

A more pervasive problem occurred when “None of the Above” was the reviewed answer. The JDI method must decide as to the best ST (unless, as rarely happens, the context is empty), hence the best disambiguator answer. Thus, when the reviewed answer for either MeSH Frequency or JDI was “None of the Above”, the disambiguator answer was always incorrect. Since neither method was designed to return “None of the Above”, it was decided to present and therefore concentrate on results ignoring those instances with this reviewed answer. Because all reviewed answers for the ambiguity *association* were “None of the Above”, this ambiguity was eliminated altogether. A side effect of ignoring “None of the Above” was to reduce the total number of instances by more than half for the ambiguities *failure*, *fit*, *lead*, *reduction*, *resistance*, and *support*, but these were included in the results anyway. One can assume that raters selected “None of the Above” for many instances of these six ambiguities on account of the fact that they are common English words corresponding to concepts not found in the Metathesaurus.

RESULTS AND ANALYSIS

Precision Analysis and Results

We ran the ambiguities comparing MeSH Frequency and the various JDI contexts. Summary precision scores and individual precision scores for the 45 ambiguities are presented in Table 1. JDI, regardless of context, performed significantly better than MeSH Frequency with average precision of .2491, versus average precision ranging from 0.7710 – 0.7873 for the JDI contexts. The median precision for MeSH Frequency was 0.0152 versus a median precision ranging from 0.8507 – 0.9048 for the JDI contexts. Twenty-two of the 45 ambiguities had 0.0000 precision score (see discussion of MeSH Frequency in the previous section for explanation) versus none for JDI.

Three of the JDI contexts (ambig-sentence, ambig-sentences, and doc-rule) approached 79% average precision; the remaining context (doc) had an average precision of 77%. The context giving the best average precision score was ambig-sentences. The doc-rule context resulted in only a slightly lower score, which is not surprising since, in the instances where there was more than one sentence containing the ambiguity, ambig-sentences was used under doc-rule as well. The ambig-sentence context scored slightly lower than doc-rule and ambig-sentences, suggesting that, on average, just the target sentence may be too little context compared to those contexts. Figure 20 is an example where a target sentence containing the ambiguity *implantation – No serious complication resulted from implantation of FOE in this series.* – resulted in the incorrect answer “Blastocyst Implantation, natural” rather than “Implantation procedure” on account of the

ST orgf having a higher score than topp for this sentence. In particular, the acronym FOE was not helpful, as in the training set it usually appears in the context of *friend or foe* and the word foe generates a higher score for orgf (which ranks 25th among the STs) than for topp (which ranks 52nd). The ambig-sentences context, which used all four sentences containing *implantation*, gave the correct answer, as did the doc context (all fourteen sentences in the citation). On average, doc scored lowest, suggesting that the entire document may be too much context compared to the others.

PLACE FIGURE 20 HERE.

PLACE TABLE 1 HERE.

The data were analyzed in terms of the number of ambiguities for which each context performed best (precision was best or tied for best), worst (precision was worst or tied for worst), or intermediate (Table 2). The contexts doc and ambig-sentence had the best precision for 21 and 20 ambiguities, respectively, and the worst precision for 22 and 18 ambiguities, respectively; these contexts performed either the best or the worst. The doc-rule context had the best performance for 20 ambiguities compared to 15 for ambig-sentences, and they were tied at 9 ambiguities for worst performance. Thus, in this analysis, it would seem that doc-rule had the edge in terms of optimum performance (balancing best and worst precision). Ignoring ambiguities where the difference between best and worst performance was less than 0.0200 (*extraction, mole, mosaic, and transient*) the data suggest that doc, which was best for 17 ambiguities and worst for 22 ambiguities, fared poorest in terms of optimum performance, while doc-rule (best for 20 ambiguities and worst for 5) remained optimally the best. Ranked second and third for optimum performance would be ambig-sentences and ambig-sentence, respectively.

PLACE TABLE 2 HERE.

We compare the optimally performing JDI method, doc rule, to two supervised methods using the WSD collection. In general, precision of JDI is comparable to these other methods. Table 3 compares JDI to the best overall naïve Bayes classifier in Leroy & Rindflesch (2004) for the thirteen ambiguities classified by both methods. For nine ambiguities, JDI precision is higher, and average JDI precision is higher. Although the Liu et al. (2004) experiment does not permit a side by side comparison, performance of all supervised classifiers (precision around 80%) on 22 of the original 50 ambiguities is comparable to that of the methods presented in Table 1.

PLACE TABLE 3 HERE.

Preliminary Performance Analysis

We have begun to analyze JDI performance failure (which we define as < 0.6500) by examining individual ambiguities. The following are some observations (refer to Appendixes I, II, and III for choices of meaning and ST) regarding poor performance:

1. Difficulty in distinguishing between chemicals and laboratory procedures. Examples include *lead* and *glucose*. In fact, the text strings “lead” and “glucose” each result in lbpr as the preferred ST, compared to elii for the former and to bacs and carb for the latter. That is, these strings have a higher association with laboratory procedures than for substance terms. Furthermore, sentences containing these words tend to have co-occurring words denoting laboratory procedures, thus boosting the lbpr score.

2. Difficulty in distinguishing between physiologic functions and their measurement or determination or the functions in terms of findings, for example *blood_pressure*, where the text has a higher association with *diap* and *lbtr* than with *orgf*.
3. Idiosyncratic Metathesaurus meanings and ST assignments, for example *pressure*, where one of the meanings is the concept Baresthesia (pressure sensation, or the physiologic discrimination of various degrees of pressure on the surface) In the ambig-sentences context, 46 of the 58 incorrect answers involved Baresthesia as the incorrect answer.
4. System's non-selection of very general ST over a very common ST, for example *fluid*, where the correct ST was *sbst* for every instance, in contrast to *qlco*, but it was selected by the system for only 3 of 67 times for the ambig-sentences context.
5. Difficulty in distinguishing between STs for two types of general activity, for example, *evaluation*, which requires distinguishing between *hlca* (the most general health care activity ST) and *resa* (research activity ST).
6. Difficulty in distinguishing between STs sharing semantic features, for example, *nutrition* which may require selecting between semantically-related STs *orga* and *orgf* as the correct ST and *japanese* requiring selecting between STs *popg* and *lang*.
7. Ambiguities where the context often does not reflect the ST of the meaning of the ambiguity. For example, human raters selected the *topp* meaning for the following ambig-sentences context for *nutrition* (where the ambiguity is the variant *nutritional*) "If women have a different metabolic response to the human immunodeficiency virus (HIV), nutritional advice may differ from HIV-seropositive man. Therefore, nutritional advice may need to vary according to the gender of the asymptomatic HIV-seropositive subject." The system's

selection for the context was orga because this was the best ST for many of the words (e.g., immunodeficiency, seropositive, HIV, virus).

For some of these poor-performance ambiguities it is also the case that the context corresponding to the meanings can be expected to be similar (i.e., have similar vocabularies) to one another. On the other hand, for several ambiguities where system performance was good (which we define as > 0.8500) the contexts corresponding to different meanings can be expected to be quite different. This difference, in turn, can be translated into contrasting STs corresponding to the words in the contexts to which JDI is sensitive. Examples of good performance include ambiguities involving:

1. natural or physiologic processes versus intentional procedures: *reduction* (npop hlca), *transport* (celf hlca), *implantation* (orgf topp)
2. laboratory versus non-laboratory environment: *determination* (gora lbpr), *culture* (idcn lbpr), *extraction* (topp lbpr)
3. temporality versus non-temporality: *transient* (popg tmco), *frequency* (tmco sosy)
4. mental versus non-mental: *inhibition* (menp moft), *resistance* (menp socb), *depression* (ftcn mobd), *condition* (qlco menp)
5. social versus non-social: *support* (socb medd), *failure* (patf socb)

FUTURE WORK

Future work falls into two categories: improving the JDI methodology and studying the use of JDI in applications.

Improving the JDI methodology (see Methodology of JDI-Based Indexing) includes updating the “ST documents” based on the latest version (2004) of the UMLS Metathesaurus. The ST documents we are using were developed in 2002. Another aspect of the methodology we will examine is the stopwords and restrictwords lists. An extensive stopword list, developed empirically, is now being used. Using JDI, we may be able to identify what constitutes a good stopword by comparing the JD vectors of generally agreed-upon stopwords with candidate stopwords. Improving the methodology includes improving its general application for solving the “None of the Above” problem. For example, if the candidate STs all score very low, is this an indication that none of them is appropriate? We also can try to adopt methods for identifying acronyms (Liu et al., 2002a; Wren & Garner, 2002; Yu et al., 2002; Schwartz & Hearst, 2003), substituting the full form for the acronym. For example, if the full form “foramen ovale electrode” had been substituted for “FOE” in the target sentence shown in Figure 20, the correct ST would have resulted. We can test changes on the WSD test collection.

Disambiguation using JDI is already being used in experimental systems at NLM, specifically in SemGen – adapted from the natural language processing (NLP) program SemRep – that identifies gene interaction predications from MEDLINE citations (Rindflesch et al., 2003; Libbus et al., 2004). JDI increases accuracy by identifying citations in the molecular genetics domain before NLP begins. JDI has also been explored for gene symbol disambiguation in connection with BITOLA, an interactive literature-based biomedical discovery support system (Hristovski et

al., in press) by being able to determine, for example, that the document title “Ethics in a twist: ‘Life Support’, BBC1” is outside the genetics domain, thereby, in effect, disambiguating the British television station BBC1, as in this title, from the symbol BBC1 for the breast basic conserved 1 gene. Based on the experiment described in the current paper, perhaps JDI can be studied further in applications necessitating WSD of strings according to various meanings associated with STs.

CONCLUSIONS

We have described an experiment using NLM’s WSD test collection to compare four versions of the Journal Descriptor Indexing methodology (based on extent of context) to a baseline MeSH Frequency methodology. For the forty-five ambiguities studied, the overall average precision of the highest scoring JDI method was 0.7873 compared to 0.2492 for MeSH Frequency. Furthermore, for the 45 individual ambiguities, average precision was > 0.90 for 23 (51%) of them, > 0.85 for 24 (53%), and > 0.65 for 35 (79%). Based on these results we feel that JDI shows promise as an unsupervised method for WSD using ready-made resources at NLM – JDs assigned to journals and thus automatically assigned to words in a large MEDLINE training set; UMLS Metathesaurus concepts assigned to STs and thus serving as ST documents (sets of words labeled by the STs). JDI uses these resources to automatically pre-label words in the training set with JDs and then with STs. Our method avoids the effort, time, and expense of hand-tagging a training set for word senses as in supervised methods. We hope to improve the performance of JDI and test its use in actual applications.

REFERENCES

Aronson A.R., Rindfleisch, T.C., & Browne, A.C. (1994). Exploiting a large thesaurus for information retrieval. In Proceedings RIAO-94 Conference (pp. 197-216). Paris: CID.

Aronson A.R., Bodenreider, O., Chang H.F., Humphrey, S.M., Mork, J.G., Nelson, S.J., Rindfleisch, T.C., & Wilbur, W.J. (2000). The NLM Indexing Initiative. Proceedings / AMIA ... Annual Symposium, 17-21.

Aronson A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA ... Annual Symposium, 17-21.

Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., & Rogers, W.J. (2004). The NLM Indexing Initiative's Medical Text Indexer. Medinfo, 11(Pt 1), 368-372.

Brown, P.F., Della Pietra, S.A., Della Pietra V.J., & Mercer, R.L. (1991), Word-sense disambiguation using statistical methods. In Proceedings of the 29th Conference of the Association for Computational Linguistics (pp. 264-270).

Clough, P., & Stevenson, M. (2004). Cross-language information retrieval using EuroWordNet and word sense ambiguity. In S. MacDonald & J. Tait (Eds.) Lecture Notes in Computer Science 2997: Advances in Information Retrieval, 26th Conference on IR Research, ECIR 2004 Proceedings (pp. 327-337). Heidelberg, Germany: Springer.

Edmonds, P., & Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8, 279-291.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hristovski, D., Peterlin, B., Mitchell, J.A., & Humphrey, S.M. (in press). Using literature-based discovery to identify disease candidates-genes. *International Journal of Medical Informatics*.

Humphrey S.M. (1998). A new approach to automatic indexing using journal descriptors. In C. M. Preston (Ed.). *Proceedings of the 61st ASIS Annual Meeting* (pp. 496-500). Medford, NJ: Information Today.

Humphrey, S.M. (1999). Automatic indexing of documents from journal descriptors: a preliminary investigation. *Journal of the American Society for Information Science*, 50, 661-674.

Humphrey, S.M., Rindfleisch, T.C., & Aronson, A.R. (2000). Automatic indexing by discipline and high-level categories: methodology and potential applications. In *Proceedings of the 11th ASIST SIG/CR Classification Research Workshop* (pp. 103-116). Silver Spring, MD: American Society for Information Science and Technology..

Ide, N., & Véronis, J. (1998). Word sense disambiguation: the state of the art. *Computational Linguistics*, 24, 1-40.

Kilgarriff, A., Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34, 15-48.

Leroy, G., & Rindflesch, T.C. (2004). Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naïve Bayes classifier. *Medinfo*, 11(Pt 1),381-385.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC)* (pp. 24-26). New York: Association for Computing Machinery.

Libbus, B., Kilicoglu, H., Rindflesch, T.C., Mork, J.G., & Aronson, A.R. (2004). Using natural language processing, LocusLink, and the Gene Ontology to compare OMIM to MEDLINE. In *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases* (pp. 69-76). East Stroudsburg, PA: Association for Computational Linguistics.

Liddy E.D., & Paik, W. (1993). From handcrafted dictionary subject codes to statistically-guided word sense disambiguation. In *Probabilistic Approaches to Natural Language, Papers from the AAAI Fall Symposium, Technical Report FS-92-05* (pp. 98-107). Menlo Park, CA: AAAI Press..

- Liddy, E.D., Paik W., & Woelfel, J.K. (1993). Use of Subject Field Codes from a machine-readable dictionary for automatic classification of documents. In *Advances in Classification Research, Proceedings of the 3rd ASIS SIG/CR Classification Workshop* (pp. 83-100). Medford, NJ: Learned Information.
- Liu H., Aronson, A.R., & Friedman, C. (2002a). A study of abbreviations in MEDLINE abstracts. *Proceedings / AMIA ... Annual Symposium*, 464-468.
- Liu H., Johnson, S.B., & Friedman, C. (2002b). Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *Journal of the American Medical Informatics Association*, 6, 621-636.
- Liu, H., Teller, V., & Friedman, C. (2004). A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association*, 11, 320-331.
- Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The role of domain knowledge in word sense disambiguation. *Natural Language Engineering*, 8, 359-373.
- Maynard, D., Ananiadou, S. (2000). Trucks: a model for automatic multiword term recognition. *Journal of Natural Language Processing*, 8, 101-126.

Mihalcea, R., & Moldovan, D.I. (1999). An automatic method for generating sense tagged corpora. In Proceedings of the Sixteenth National Conference on Artificial Intelligence, Eleventh Conference on Innovative Applications of Artificial Intelligence (pp. 461-466). Menlo Park, CA: American Association for Artificial Intelligence.

Mihalcea, R., Chklovsky, T., & Kilgarriff, A. (2004). The SENSEVAL-3 English lexical sample task. In Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for The Semantic Analysis of Text [CD-ROM] (pp. 25-28). New Brunswick, NJ: Association for Computational Linguistics..

NLM (2002). List of journals indexed in Index Medicus 2002. NIH Publication No. 02-267. Bethesda, MD: National Library of Medicine.

NLM Indexing Initiative. Retrieved on May 19, 2004, from <http://ii.nlm.nih.gov/>

NLM Medical Subject Headings. Retrieved on May 19, 2004, from <http://www.nlm.nih.gov/mesh/2004/MBrowser.html>.

NLM Unified Medical Language System. Retrieved on May 19, 2004, from <http://nlm.nih.gov/research/umls/>

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In Conference on Knowledge Discovery in Data, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 613-619). New York: ACM Press.

Pasca, M., & Harabagiu, S. (2001). The informative role of WordNet in open-domain question answering. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01), Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (pp. 138-143). East Stroudsburg, PA: Association for Computational Linguistics.

Pedersen, T., & Bruce, R. (1997). Distinguishing word senses in untagged text. In C. Cardie & R. Weischedel (Eds.), Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2). Somerset, NJ: ACL.

Resnik, P. (2004). Exploiting hidden meanings: using bilingual text for monolingual annotation. In A. Gelbukh (Ed.), Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Proceedings (pp. 283-299), Heidelberg, Germany: Springer.

Rindfleisch, T.C., Libbus, B., Hristovski, D., Aronson, A.R., & Kilicoglu, H. (2003). Semantic relations asserting the etiology of genetic diseases. AMIA ... Annual Symposium proceedings [electronic resource] / AMIA Symposium, 554-558.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval (p. 124). New York: McGraw-Hill, 1983.

Schütze H. (1992). Dimensions of meaning. In Proceedings Supercomputing '92 (pp. 787-796). Los Alamitos, CA: IEEE Comput Soc Press.

Schwartz, A.S., & Hearst, M.A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. Pacific Symposium on Biocomputing, 451-62.

Strapparava, C., Giuliano, C., & GlioZZo, A. (2004). Pattern abstraction and term similarity for word sense disambiguation: IRST at SENSEVAL-3. In Proceedings of SENSEVAL-3: The Third International Workshop on the Evaluation of Systems for The Semantic Analysis of Text [CD-ROM] (pp. 229-234). New Brunswick, NJ: Association for Computational Linguistics.

Vorhees, E. (1998). Using WordNet for text retrieval. In C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database (pp. 285-303). Cambridge, MA: MIT Press.

Weeber, M., Mork, J.G., & Aronson, A.R. (2001). Developing a test collection for biomedical word sense disambiguation. Proceedings / AMIA ... Annual Symposium, 746-750.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., & Buitelaar, P. (2003).

Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using

UMLS. In Natural Language Processing in Biomedicine ACL 2003 Workshop (pp. 9-16). East Stroudsburg, PA: Association for Computational Linguistics.

Wren, J.D., & Garner, H.R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*, 41, 426-434.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings the 14th International Conference on Computational Linguistics (COLING-92) (pp. 454-460). International Committee on Computational Linguistics. East Stroudsburg, PA: Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In 33rd Annual Meeting of the Association for Computational Linguistics, Proceedings of Conference (pp. 189-196). San Francisco: Morgan Kaufmann,

Yu, H., Hripcsak, G., & Friedman, C. (2002). Mapping abbreviations to full forms in electronic articles. *Journal of the American Medical Informatics Association*, 9, 262-272.

Appendix I. WSD test collection ambiguities and respective STs and Metathesaurus concepts.

adjustment	ftcn "Adjustment Action"; inbe "Individual Adjustment"; menp "Psychological Adjustment"
association	menp "Mental association"; soch "Relationship by association"
blood_pressure	lbtr "Arterial pressure"; orgf "Blood Pressure <1>"; diap "Blood Pressure Determination"
cold	dsyn "Common Cold" "Chronic Obstructive Airway Disease"; qlco "Cold Sensation"; soty "Cold Sensation"; topp "Cold Therapy"; npop "cold temperature"
condition	qlco "Condition"; menp "Conditioning (Psychology)"
culture	idcn "Anthropological Culture"; lbpr "Laboratory culture"
degree	qlco "degree <1>"; inpr "degree <2>"
depression	ftcn "Depression motion"; mobd "Mental Depression"
determination	gora "adjudication"; lbpr "determination <2>"
discharge	bdsu "Discharge, Body Substance"; hlca "Patient Discharge"
energy	npop "Energy (physics)"; fndg "Vitality"
evaluation	inpr "Evaluation"; resa "Evaluation"; hlca "Health evaluation"
extraction	topp "Extraction, NOS"; lbpr "extraction <1>"
failure	patf "Failure, NOS"; soch "failure <1>"
fat	lipd "Fatty acid glycerol esters"; orga "Obese build"
fit	fndg "Fit and well"; dsyn "Siezures"; soty "Siezures"
fluid	qlco "Fluid <2>"; sbst "Liquid substance, NOS"
frequency	tmco "Frequencies"; soty "Increased frequency of micturation"
ganglion	acab "Benign cystic mucinous tumour"; bpoc "Ganglia"
glucose	bacs "Glucose"; carb "Glucose"; lbpr "Glucose measurement"
growth	orgf "Growth <1>"; ftcn "growth <2>"
immunosuppression	orgf "Natural immunosuppression"; topp "Therapeutic immunosuppression"
implantation	topp "Blastocyst Implantation, natural"; topp "Implantation procedure"
inhibition	menp "Psychological inhibition"; moft "inhibition, physical"
japanese	popg "Japanes Population"; lang "Japanese language"
lead	elii "Lead"; lbpr "Lead measurement, quantitative"
man	humn "Homo sapiens"; popg "Men" "Homo sapiens"; orga "Male"
mole	neop "Benign melanocytic nevus of skin"; mamm "Mole the mammal"; qnco "mol"
mosaic	inpr "Mosaic <4>"; orga "Mosaicism <1>"; spco "Spatial Mosaic"
nutrition	topp "Feeding and dietary regimes"; orga "Nutrition"; bmod "Science of nutrition"; orgf "Science of nutrition"
pathology	bmod "Pathology"; patf "pathology <3>"
pressure	ortf "Baresthesia"; topp "Pressure - action"; qnco "Pressure- physical agent"
radiation	npop "Electromagnetic Energy"; topp "Radiation therapy"
reduction	npop "Reduction (chemical)"; hlca "Reduction - action"
repair	topp "Repair - action"; orgf "Wound Healing"
resistance	menp "Resistance <2>"; soch "resistance <1>"
scale	bpoc "Integumentary scale"; inpr "Intellectual scale"; mnob "Weight measurement scales"
secretion	bdsu "Bodily secretions"; biof "secretion <3>"
sensitivity	lbtr "Antimicrobial susceptibility"; fndg "Personality sensitivity"; menp "Personality sensitivity"; qnco "Statistical sensitivity"
sex	inbe "Coitus"; orgf "Coitus"; orga "Gender" "Sex <2>"
single	qnco "Singular"; popg "Unmarried <2>"
strains	inpr "Microbiology subtype strains"; inpo "Muscle strain"
support	soch "Support"; medd "Support, NOS"
surgery	topp "Surgery <3>"; bmod "Surgery specialty"
transient	popg "Transient Population Group"; tmco "Transitory"
transport	celf "Biological Transport"; hlca "Patient Transport"
ultrasound	npop "Ultrasonic Shockwave"; diap "Ultrasonography"
variation	qlco "Variant"; npop "Variation (Genetics)"
weight	orga "Body Weight"; qnco "Body Weight" "Weight";
white	popg "Caucasoid Race"; qlco "White color"

Appendix II. ST abbreviations and corresponding full forms represented in the WSD test collection.

acab Acquired Abnormality
 bacs Biologically Active Substance
 bdsu Body Substance
 biof Biologic Function
 bmod Biomedical Occupation or Discipline
 bpoc Body Part, Organ, or Organ Component
 carb Carbohydrate
 celf Cell Function
 diap Diagnostic Procedure
 dsyn Disease or Syndrome
 findg Finding
 fcn Functional Concept
 gora Government or Regulatory Activity
 hlca Health Care Activity
 humn Human
 idcn Idea or Concept
 inbe Individual Behavior
 inpr Intellectual Product
 lang Language
 lbpr Laboratory Procedure
 lbtr Laboratory or Test Result
 lipd Lipid
 mamm Mammal
 medd Medical Device
 menp Mental Process
 mnob Manufactured Object
 mobd Mental or Behavioral Dysfunction
 moft Molecular Function
 neop Neoplastic Process
 npop Natural Phenomenon or Process
 orga Organism Attribute
 orgf Organism Function
 ortf Organ or Tissue Function
 patf Pathologic Function
 popg Population Group
 qlco Qualitative Concept
 qnco Quantitative Concept
 resa Research Activity
 sbst Substance
 socb Social Behavior
 sosy Sign or Symptom
 spco Spatial Concept
 tmco Temporal Concept
 topp Therapeutic or Preventive Procedure

Appendix III. Hierarchical view of ST abbreviations and corresponding full forms represented in the WSD test collection.

Event	Activity	Behavior	soch Social Behavior inbe Individual Behavior
		Occupational Activity	hlca Health Care Activity lbpr Laboratory Procedure diag Diagnostic Procedure topp Therapeutic or Preventive Procedure resa Research Activity gora Government or Regulatory Activity
	Phenomenon or Process	npop Natural Phenomenon or Process biof Biologic Function	Physiologic Function orgf Organism Function menp Mental Process ortf Organ or Tissue Function celf Cell Function mofl Molecular Function patf Pathologic Function dsyn Disease or Syndrome mobd Mental or Behavioral Dysfunction neop Neoplastic Process
Entity	Physical Object	Organism	Animal mamm Mammal humn Human
		Anatomical Structure	Fully Formed Anatomical Structure bpoc Body Part, Organ, or Organ Component Anatomical Abnormality acab Acquired Abnormality
		mnob Manufactured Object medd Medical Device	
		sbst Substance bdsu Body Substance Chemical	Chemical Viewed Structurally Organic Chemical carb Carbohydrate lipd Lipid elii Element, Ion, or Isotope Chemical Viewed Functionally bacs Biologically Active Substance
	Conceptual Entity	orga Organism Attribute findg Finding	lbtr Laboratory or Test Result sosy Sign or Symptoms
		idcn Idea or Concept tmco Temporal Concept qlco Qualitative Concept qnco Quantitative Concept spco Spatial Concept ftcn Functional Concept	
		Occupation or Discipline bmod Biomedical Occupation or Discipline	
		Group popg Population Group inpr Intellectual Product lang Language	

(002) -- UI - 98008895 TI - Expression of betacellulin and epiregulin genes in the mouse uterus temporally by the blastocyst solely at the site of its apposition is coincident with the "window" of implantation. AB - In the mouse, the process of implantation is initiated by the attachment reaction between the blastocyst trophoctoderm and uterine luminal epithelium that occurs at 2200-2300 h on day 4 (day 1 = vaginal plug) of pregnancy. Several members of the EGF family are considered important in embryo-uterine interactions during implantation. This investigation demonstrates that the expression of two additions to the family, betacellulin and epiregulin, are exquisitely restricted to the mouse uterine luminal epithelium and underlying stroma adjacent to the implanting blastocyst. These genes are not expressed during progesterone-maintained delayed implantation, but are rapidly switched on in the uterus surrounding the implanting blastocyst following termination of the delay by estrogen. These results provide evidence that expression of betacellulin and epiregulin in the uterus requires the presence of an active blastocyst and suggest an involvement of these growth factors in the process of implantation. Copyright 1997 Academic Press.

Choices	M1	M2	None	Result
Counts	8	0	0	M1

M1 - Implantation <1> (Blastocyst Implantation, natural) [orgf, Organism Function]
M2 - Implantation <2> (Implantation procedure) [topp, Therapeutic or Preventive Procedure]
None - None of the Above

Figure 1. Result of choices of eight raters who used the WSD interface to disambiguate s1, unanimously selecting “Blastocyst Implantation, natural” (having ST orgf).

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
1	orgf	Organism Function	0.5897
14	spco	Spatial Concept	0.4841
15	diap	Diagnostic Procedure	0.4831
18	topp	Therapeutic or Preventive Procedure	0.4591
25	emst	Embryonic Structure	0.4301
41	aapp	Amino Acid, Peptide, or Protein	0.3724
104	vtbt	Vertebrate	0.2210

Figure 2. ST indexing of s1 *“In the mouse, the process of implantation is initiated by the attachment reaction between the blastocyst trophoctoderm and uterine luminal epithelium that occurs at 2200-2300 h on day 4 (day 1 = vaginal plug) of pregnancy.”*

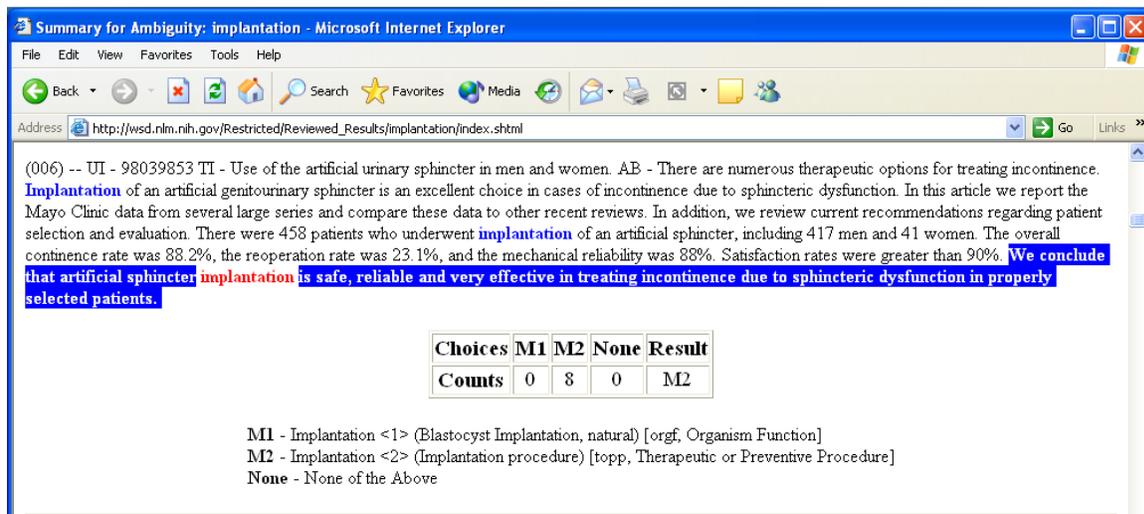


Figure 3. Result of choices of eight raters who used the WSD interface to disambiguate s1, unanimously selecting “Implantation procedure” (having ST topp).

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
1	diap	Diagnostic Procedure	0.6238
2	topp	Therapeutic or Preventive Procedure	0.6098
3	spco	Spatial Concept	0.5627
9	orgf	Organism Function	0.4797
59	aapp	Amino Acid, Peptide, or Protein	0.2739
85	emst	Embryonic Structure	0.2181
119	vtbt	Vertebrate	0.1349

Figure 4. ST indexing of s2 “*We conclude that artificial sphincter implantation is safe, reliable and very effective in treating incontinence due to sphincteric dysfunction in properly selected patients.*”

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
57	aapp	Amino Acid, Peptide, or Protein	0.3373
5	diap	Diagnostic Procedure	0.6637
39	emst	Embryonic Structure	0.4168
13	orgf	Organism Function	0.6013
1	spco	Spatial Concept	0.7027
2	topp	Therapeutic or Preventive Procedure	0.6937
108	vtbt	Vertebrate	0.1748

Figure 5. Items in ST vector for *implantation*.

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
24	aapp	Amino Acid, Peptide, or Protein	0.2160
44	diap	Diagnostic Procedure	0.1728
1	emst	Embryonic Structure	0.6096
2	orgf	Organism Function	0.4998
46	spco	Spatial Concept	0.1654
45	topp	Therapeutic or Preventive Procedure	0.1695
41	vtbt	Vertebrate	0.1780

Figure 6. Items in ST vector for *blastocyst*.

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
66	aapp	Amino Acid, Peptide, or Protein	0.1638
1	diap	Diagnostic Procedure	0.6746
100	emst	Embryonic Structure	0.1068
21	orgf	Organism Function	0.3584
3	spco	Spatial Concept	0.5660
2	topp	Therapeutic or Preventive Procedure	0.6528
118	vtbt	Vertebrate	0.0518

Figure 7. Items in ST vector for *sphincter*.

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
1	orgf	Organism Function	0.5506
4	emst	Embryonic Structure	0.5132
12	spco	Spatial Concept	0.4340
13	topp	Therapeutic or Preventive Procedure	0.4316
16	diap	Diagnostic Procedure	0.4182
45	aapp	Amino Acid, Peptide, or Protein	0.2766
92	vtbt	Vertebrate	0.1764

Figure 8. ST indexing of *blastocyst implantation*.

<u>Rank</u>	<u>ST abbr</u>	<u>ST</u>	<u>Score</u>
1	topp	Therapeutic or Preventive Procedure	0.6732
2	diap	Diagnostic Procedure	0.6692
3	spco	Spatial Concept	0.6344
18	orgf	Organism Function	0.4798
59	emst	Embryonic Structure	0.2618
62	aapp	Amino Acid, Peptide, or Protein	0.2506
116	vtbt	Vertebrate	0.1133

Figure 9. ST indexing of *sphincter implantation*.

JID	0372772
TI	Fertility and Sterility
TA	Fertil Steril
JD	Reproduction

Figure 10. NLM's journal record for *Fertility and Sterility* showing the JD Reproduction.

PMID 10856474
TI Blastocyst score affects *implantation* and pregnancy outcome: towards a single blastocyst transfer.
JID 0372772
SO Fertil Steril 2000 Jun;73(6):1155-8.
*JD Reproduction

*mapped from the journal record for Fertil Steril (Figure 10).

Figure 11. Sample MEDLINE citation in the training set showing inheritance of JD from NLM's journal record.

<u>Rank</u>	<u>JD</u>	<u>Score</u>
109	Acquired Immunodeficiency Syndrome	0.0000
4	Biomedical Engineering	0.4067
2	Cardiology	0.6416
3	Ophthalmology	0.6405
5	Otolaryngology	0.3741
1	Reproduction	0.9044
109	Zoology	0.0000

Figure 12. Items in JD vector for *implantation*.

<u>Rank</u>	<u>Score</u>	<u>JD</u>
1	0.1431	Reproduction
2	0.0747	Obstetrics
3	0.0735	Gynecology
4	0.0257	Embryology
5	0.0245	Veterinary Medicine

Figure 13. JD indexing of s1 *“In the mouse, the process of implantation is initiated by the attachment reaction between the blastocyst trophoctoderm and uterine luminal epithelium that occurs at 2200-2300 h on day 4 (day 1 = vaginal plug) of pregnancy.”*

<u>Rank</u>	<u>Score</u>	<u>JD</u>
1	0.1857	Urology
2	0.0522	Gynecology
3	0.0504	Gastroenterology
4	0.0423	Obstetrics
5	0.0321	Reproduction

Figure 14. JD indexing of s2 “*We conclude that artificial sphincter implantation is safe, reliable and very effective in treating incontinence due to sphincteric dysfunction in properly selected patients.*”

<u>Rank</u>	<u>JD</u>	<u>Score</u>
83	Acquired Immunodeficiency Syndrome	0.0213
4	Ophthalmology	0.3160
5	Orthopedics	0.3070
1	Otolaryngology	0.4827
3	Surgery	0.4740
2	Urology	0.4803
127	Zoology	0.0000

Figure 15. Items in JD vector for topp (Therapeutic or Preventive Procedure) document

(arthroplasty, bandaging, dissection, hemodialysis, immunization ...).

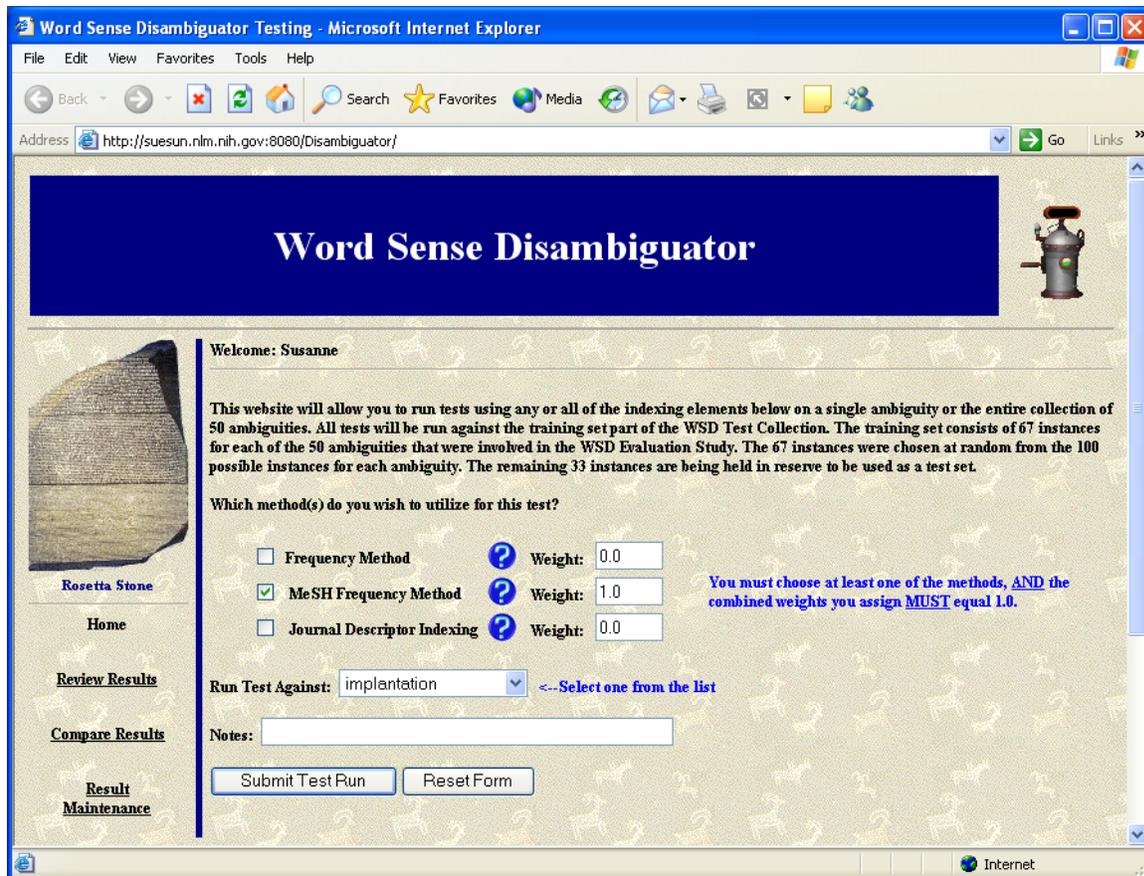


Figure 16. Word Sense Disambiguator interface where the indexing method (e.g., MeSH Frequency Method) and ambiguities, e.g., *implantation*, are selected

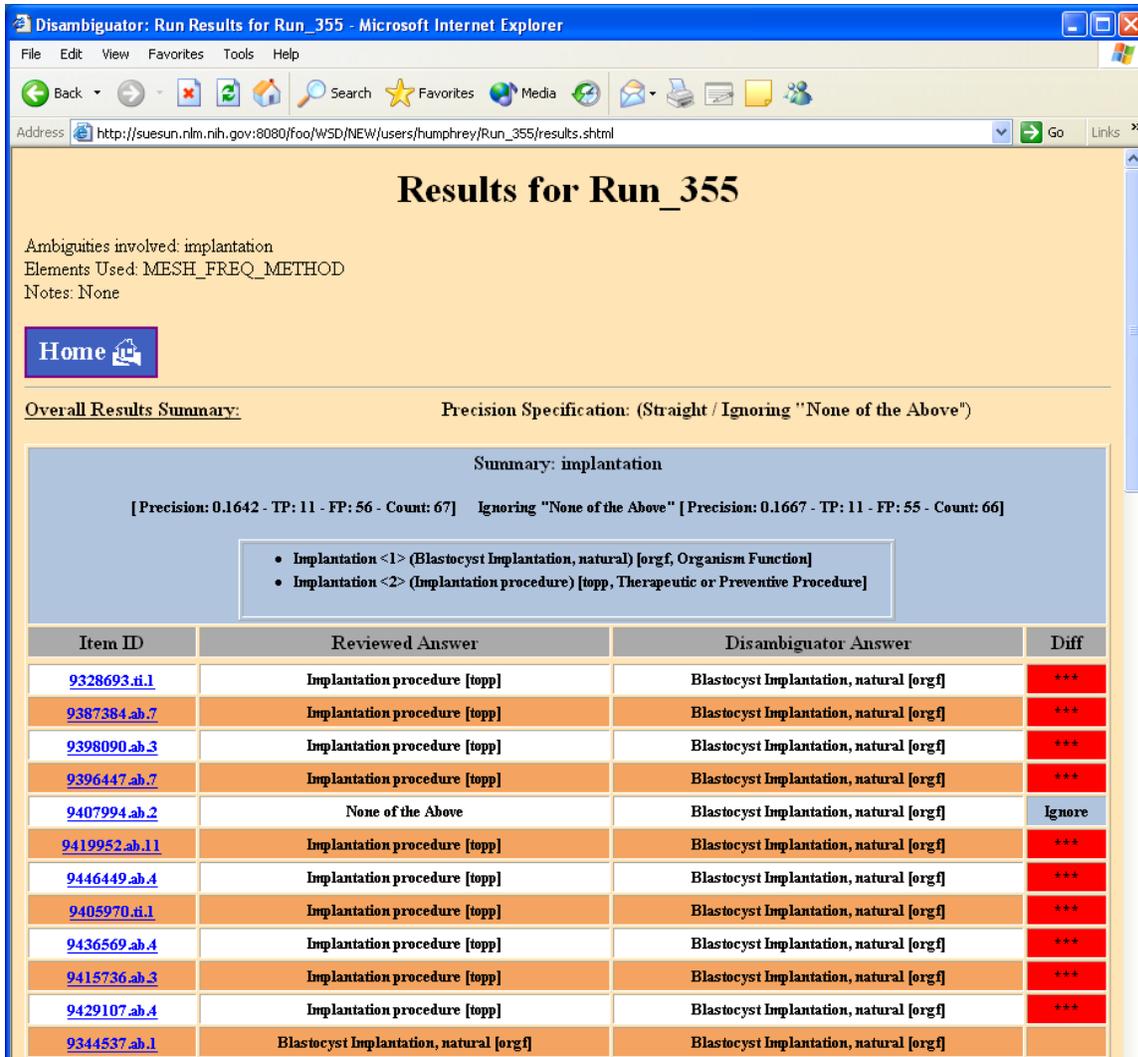


Figure 17. Word Sense Disambiguator display for MeSH Frequency results for *implantation* ambiguity, where “Blastocyst Implantation, natural” is the Disambiguator answer for all 67 instances.

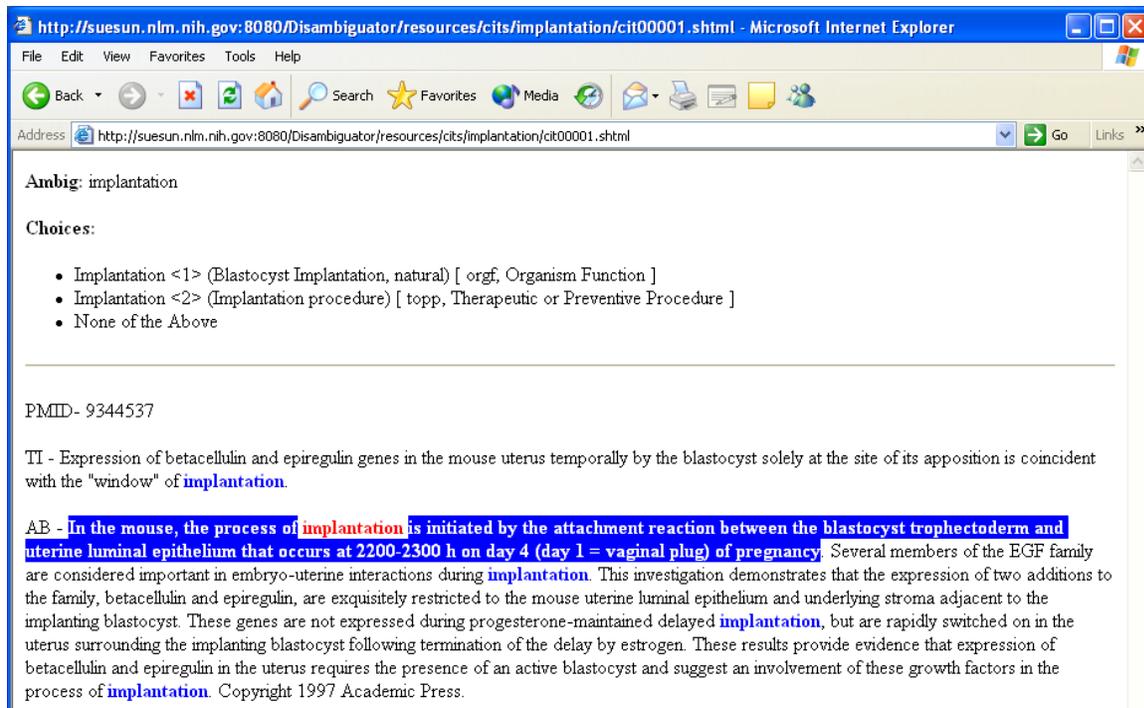


Figure 18. Word Sense Disambiguator display for MeSH Frequency results for particular *implantation* ambiguity item corresponding to s1.

Context name	Description
ambig-sentence	The one sentence containing the ambiguous string reviewed by raters (which we call the <i>target sentence</i>)
doc	The entire citation
ambig-sentences	All sentences containing the ambiguous string or its variants
doc-rule	If ambig-sentence = ambig-sentences and ambig-sentence has fewer words than some threshold, then use doc

Figure 19. Contexts for ambiguous instances.

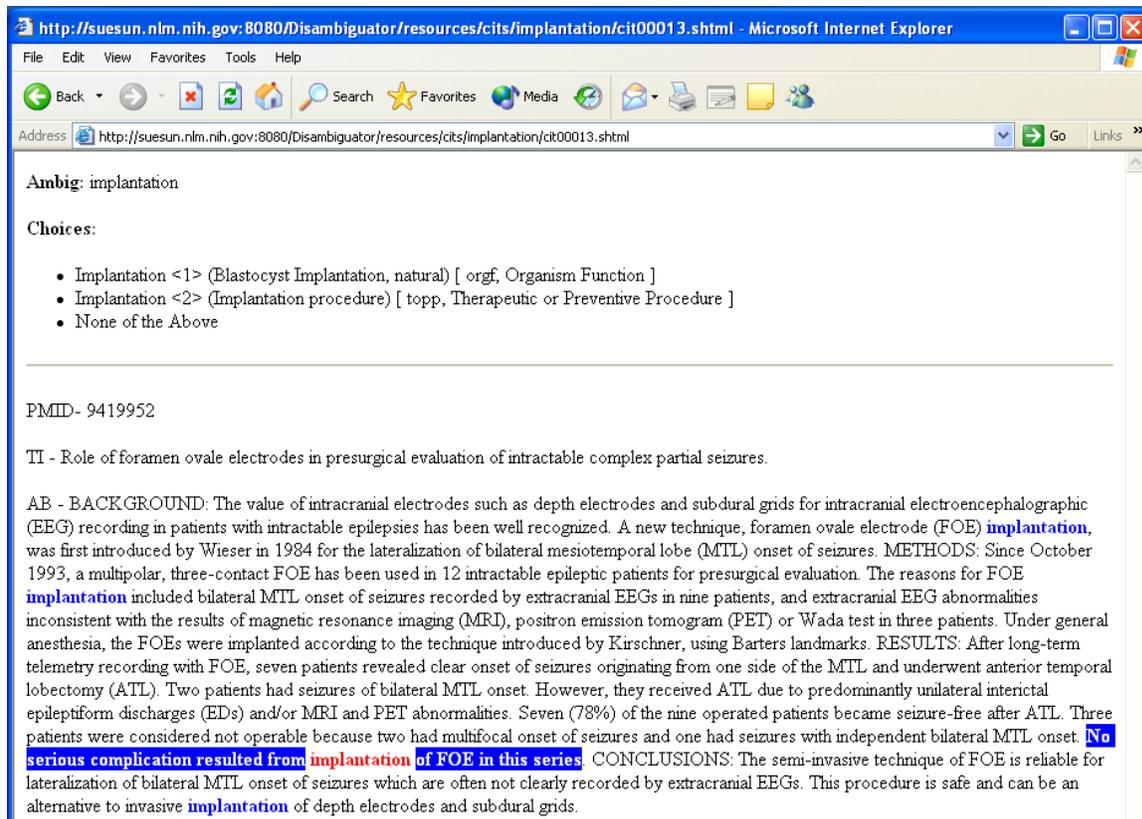


Figure 20. Example of target sentence with too little context including the acronym FOE which contributes to the wrong answer.

Ambiguities	MeSH Frequency precision	JDI doc context precision	JDI ambig- sentence context precision	JDI ambig- sentences context precision	JDI doc-rule context precision	number of instances
Summary						
average	0.2492	0.7710	0.7860	0.7873	0.7870	54
median	0.0152	0.8507	0.8939	0.9048	0.9048	63
range	0.0000 – 1.0000	0.0448 – 1.0000	0.0448 – 1.0000	0.0448 – 1.0000	0.0597 – 1.0000	3 – 67
Individual						
adjustment	0.1000	0.8167	0.6333	0.7500	0.7667	60
blood_pressure	0.0000	0.4030	0.4478	0.4179	0.4179	67
condition	0.0169	0.8983	0.9322	0.9322	0.9322	59
culture	0.1045	1.0000	0.9552	0.9851	1.0000	67
degree	0.0000	0.9318	0.9545	0.9545	0.9773	44
depression	1.0000	0.8070	0.9474	0.9474	0.9474	57
determination	0.0000	1.0000	1.0000	1.0000	1.0000	54
discharge	1.0000	0.8889	0.9630	0.9630	0.9259	54
energy	0.0000	0.6418	0.8358	0.7313	0.7015	67
evaluation	0.0000	0.5522	0.5672	0.5821	0.5970	67
extraction	0.0000	1.0000	0.9831	0.9831	0.9831	59
failure	0.0000	1.0000	0.9444	0.9444	0.9444	18
fat	0.9583	0.6250	0.7917	0.7500	0.7500	48
fit	0.0000	1.0000	1.0000	1.0000	1.0000	12
fluid	0.0000	0.0448	0.0448	0.0448	0.0597	67
frequency	0.0000	0.8889	0.9683	0.9048	0.9048	63
ganglion	0.9403	0.9403	0.9403	0.9403	0.9403	67
glucose	0.9254	0.4179	0.3582	0.3881	0.3881	67
growth	0.0000	0.7463	0.6567	0.7015	0.7015	67
immunosuppression	0.5224	0.6866	0.6866	0.7612	0.7463	67
implantation	0.1667	0.8939	0.8939	0.9242	0.9394	66
inhibition	0.0000	0.9851	0.9254	1.0000	0.9851	67
japanese	0.0000	0.4717	0.5849	0.5660	0.5472	53
lead	0.8889	0.2778	0.3889	0.3889	0.3889	18
mole	0.0182	1.0000	0.9818	0.9818	0.9818	55
mosaic	0.0000	0.6923	0.6769	0.6769	0.6769	65
nutrition	0.1774	0.4032	0.3871	0.3871	0.3548	62
pathology	0.1493	0.7164	0.7463	0.7463	0.7463	67
pressure	1.0000	0.1364	0.1061	0.1212	0.1212	66
radiation	0.4242	0.8030	0.7576	0.8030	0.7879	66
reduction	0.0000	1.0000	1.0000	1.0000	1.0000	10
repair	0.2727	0.9318	0.8636	0.8636	0.8636	44
resistance	0.0000	1.0000	1.0000	1.0000	1.0000	3
scale	0.0000	0.5116	0.7209	0.6279	0.6047	43
secretion	0.0149	0.9104	0.9403	0.9403	0.9403	67
sensitivity	0.0000	0.8286	0.8857	0.8286	0.8286	35
single	0.0000	0.9701	0.9851	0.9851	1.0000	67
strains	0.0000	0.9516	0.9677	0.9839	0.9839	62
support	0.0000	1.0000	1.0000	1.0000	1.0000	7
surgery	0.0149	0.8507	0.9851	0.9254	0.9254	67
transient	0.0000	1.0000	1.0000	0.9851	0.9851	67
transport	0.9844	1.0000	0.9531	0.9688	0.9844	64
ultrasound	0.8209	0.8060	0.8507	0.8060	0.8060	67
variation	0.1791	0.7164	0.6567	0.7015	0.7313	67
white	0.5333	0.5500	0.5000	0.5333	0.5500	60

Table 1. Summary and individual precision scores comparing MeSH Frequency disambiguation and JDI disambiguation for four contexts studied (doc, ambig-sentence, ambig-sentences, and doc-rule).

Context	No. of ambiguities best precision		No. of ambiguities worst precision		No. of ambiguities intermediate precision		Total No. of ambiguities	
		*		*		*		*
doc	21	17 *	22	22 *	2	2 *	45	41 *
ambig-sentence	22	21 *	18	15 *	5	5 *	45	41 *
ambig-sentences	15	15 *	9	5 *	21	21 *	45	41 *
doc-rule	20	20 *	9	5 *	16	16 *	45	41 *

* ignoring ambiguities *extraction*, *mole*, *mosaic*, and *transient*, where the difference between worst and best precision was < 0.0200 .

Table 2. Comparison of JDI contexts in terms of number of ambiguities where precision was best, worst, and intermediate, suggesting optimum performance.

Ambiguities	JDI precision	naïve Bayes precision
adjustment	.7667	.57
blood_pressure	.4179	.46
degree	.9773	.68
evaluation	.5970	.57
growth	.7015	.62
immunosuppression	.7463	.63
mosaic	.6769	.66
nutrition	.3548	.48
radiation	.7879	.72
repair	.8636	.81
scale	.6047	.84
sensitivity	.8286	.70
white	.5500	.62
Average	.6826	.64

Table 3. Comparison of best overall JDI disambiguation method and naïve Bayes classifier method.