Journal of Biomedical Informatics 46 (2013) 1116–1124

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text

Bridget T. McInnes [a,*], Ted Pedersen [b]

[a] Minnesota Supercomputing Institute, University of Minnesota, 117 Pleasant St SE, Minneapolis, MN 55455
[b] Department of Computer Science, University of Minnesota, 1114 Kirby Drive, Duluth, MN 55812, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Accepted 17 August 2013
Available online 4 September 2013

Keywords:
Natural language processing
NLP
Word sense disambiguation
WSD
Semantic similarity and relatedness
Biomedical documents

Introduction: In this article, we evaluate a knowledge-based word sense disambiguation method that determines the intended concept associated with an ambiguous word in biomedical text using semantic similarity and relatedness measures. These measures quantify the degree of similarity or relatedness between concepts in the Unified Medical Language System (UMLS). The objective of this work is to develop a method that can disambiguate terms in biomedical text by exploiting similarity and relatedness information extracted from biomedical resources and to evaluate the efficacy of these measure on WSD.

Method: We evaluate our method on a biomedical dataset (MSH-WSD) that contains 203 ambiguous terms and acronyms.

Results: We show that information content-based measures derived from either a corpus or taxonomy obtain a higher disambiguation accuracy than path-based measures or relatedness measures on the MSH-WSD dataset.

Availability: The WSD system is open source and freely available from http://search.cpan.org/dist/UMLS-SenseRelate/. The MSH-WSD dataset is available from the National Library of Medicine http://wsd.nlm.nih.gov.

1. Introduction

Word Sense Disambiguation (WSD) is the task of automatically identifying the intended sense (or concept) of an ambiguous word based on the context in which the word is used. In our work, the set of possible meanings for a word are defined by Concept Unique Identifiers (CUIs) associated with a particular term in the Unified Medical Language System (UMLS). Thus, when performing WSD of biomedical terms, our more specific goal is to assign a term one of its possible CUIs based on its surrounding context. For example, the term *cold* could refer to the temperature (C0009264) or the common cold (C0009443), depending on the context in which it occurs.

Automatically identifying the intended concept of ambiguous words improves the performance of clinical and biomedical applications such as medical coding and indexing for quality assessment, cohort discovery and other secondary uses of data. These capabilities are becoming essential tasks due to the growing amount of information available to researchers, the transition of health care documentation towards electronic health records, and the push for quality and efficiency in health care.

The SenseRelate algorithm introduce by Patwardhan et al. [1] determines the most context-appropriate concept of an ambiguous word using the degree of semantic similarity or relatedness between the possible concepts and the terms surrounding the ambiguous word. The underlying assumption of the algorithm is that an ambiguous word will refer to the concept that is most similar to the concepts associated with the terms that surround it.

We classify semantic similarity measures in three categories: path-based measures which rely on the hierarchical relations between the terms in a taxonomy; corpus-based information content (IC) measures which augment the path information with probabilities derived from a corpus of text; and taxonomy-based IC measures which calculate the information content of a concept based on its specificity within a taxonomy. Relatedness measures use the terms found in the definitions of concepts and possibly augment those definitions with information derived from corpora. One such measure, *vector*, uses secondary co-occurrence information obtained from a corpus to determine the relatedness between terms.

In this article, we compare path-based similarity measures, corpus-based and taxonomy-based information content similarity measures, and relatedness measures. Previous studies compared

* Corresponding author.
E-mail addresses: btmcinnes@gmail.com (B.T. McInnes), tpederse@d.umn.edu (T. Pedersen).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.008

just path and corpus-based information [2] or path-based and taxonomy-based measures [3]. Overall, we found the corpus-based similarity measures perform on par or better than the taxonomy-based measures and significantly better than the path-based and relatedness measures for the task of WSD.

Section 4 describes the resources used in this work. Section 2 describes previous knowledge-based WSD methods. Section 3 describes the semantic similarity and relatedness measures used in this work. Section 5 describes our method. The data used to evaluate the method is described in Section 6. The experiments are described in Section 7 and their results in Section 8, and a comparison to previous work in Section 9. Finally, conclusions and future work are presented in Section 10.

2. Related work

Existing methods that have been proposed to automatically disambiguate words in text can be classified into three groups: supervised [4,5], unsupervised [6,7] and knowledge-based methods [8].

Supervised methods use machine learning algorithms to assign concepts to instances containing the ambiguous word. The disadvantage of these types of methods is that training data needs to be created for each target word to be disambiguated. Whether this is done manually or automatically, it is infeasible to create such data on a large scale. Knowledge-based methods do not use manually or automatically generated training data, but use information from an external knowledge source and possibly a corpus of text. Unsupervised methods use the distributional characteristics of an outside corpus and do not rely on concept information or a knowledge source. In this work, we focus on knowledge-based methods.

In the biomedical domain, Humphrey et al. [9] introduce a knowledge-based method that assigns a concept to a target word by first identifying its semantic type with the assumption that each possible concept has a distinct semantic type. A semantic type (st-) vector is created for the semantic type of each possible concept using one word terms in the UMLS that have been assigned that semantic type. A target word (tw-) vector is created using the words surrounding the target word. The cosine of the angle between the tw-vector and each of the st-vectors is calculated and the concept whose st-vector is closest to the tw-vector is assigned to the target word. The limitation of this method is that two possible concepts may have the same semantic type. For example, the term *cortices* can refer to either the cerebral cortex (C0007776) or the kidney cortex (C0022655); each with the semantic type "Body Part, Organ, or Organ Component". Analysis of the 2009 Medline data[1] shows that there are 1,072,902 terms in Medline that exist in the UMLS of which 35,013 are ambiguous and 2979 have two or more concepts with the same semantic type. This indicates that approximately 12% of the ambiguous words cannot be disambiguated using this method.

Alexopoulou et al. [10] introduce the "Closest Sense" method which calculates the average shortest distance between the semantic type of a possible concept and the semantic types each of the words surrounding the target word. This is done for each possible concept, and the concept with the shortest distance is assigned to the target word. This method also assumes that each possible concept has a distinct semantic type.

Jimeno-Yepes et al. [11] introduce a variation of the MRD method which can be seen as a variation of the Lesk algorithm [12]. In this method, a concept vector (c-vector) for each possible concept of a target word is created using the definition information from the UMLS. A target word (tw-) vector is created using the words surrounding the target word. The cosine of the angle between the tw-vector and each of the c-vectors is calculated and the concept whose c-vector is closest to the tw-vector is assigned to the target word. Rather than the vectors containing frequency scores, the frequency of the terms in the vector are normalized based on their inverted concept frequency so that terms which are repeated many times within the UMLS will have less relevance. The results of subsequent experiments conducted by Jimeno-Yepes et al. [13] compared with those conducted previously by McInnes [14] show that the inverted concept frequency significantly increases the disambiguation accuracy of the MRD method.

Jimeno-Yepes et al. [11] also introduce the AEC method, a semi-supervised approach where instances of target word are trained on automatically generate training data from Medline. Medline is manually indexed with Medical Subject Headings (MeSH) terms where each term has an associated CUI in the UMLS. Citations from Medline that contain the target word and have been annotated with one of the possible senses of the target word are extracted. These citations are used as training data into a supervised WSD algorithm. Their results show that the AEC method obtained a higher disambiguation accuracy than MRD method discussed above and the PageRank method introduced by Agirre et al. [15].

Stevenson et al. [16] introduce a modification of the PageRank algorithm called Personalized Page Rank adapted by Agirre et al. [15] for WSD. PageRank is technique for scoring the vertices according to their importance in the overall structure of a graph. In this method, a vector is constructed containing the concepts of the context words surrounding the target word. PageRank is then applied over this subgraph and the concept in the graph with maximal score is assigned to the target word. The results show that Personalized PageRank obtains a higher disambiguation accuracy than PageRank and on par with the AEC method.

Garla and Brandt [3] use the SenseRelate algorithm proposed by Patwardhan et al. [1] to evaluate path-based and taxonomy-based similarity measures. In this method, each possible concept of an ambiguous word is assigned a score by summing the similarity score between it and the terms surrounding it. The authors also evaluate obtaining the surrounding concepts using cTAKES and MetaMap finding that MetaMap performs best on biomedical text where cTAKES performs best on clinical. The results show that for biomedical text the measure taxonomy-based information content measure obtained a higher disambiguation accuracy than the path-based measures, but on clinical text the reverse was found.

3. Similarity and relatedness measures

Relatedness measures quantify the degree to which two words are associated with each other (*scissors-paper*). Similarity is a subset of relatedness and quantifies how alike two concepts are based on their location within an *is-a* hierarchy (*car-vehicle*). This section describes the similarity and relatedness measures used in this work.

3.1. Similarity measures

Existing semantic similarity measures can be categorized into two groups: path-based and information content (IC)-based. Path-based measures rely on the shortest path information, whereas IC-based measures incorporate the probability of the concept occurring in a corpus of text.

3.1.1. Path-based

Rada et al. [17] introduce the conceptual distance measure which is the length of the shortest path between two concepts ($c1$ and $c2$) in MeSH using RB/RN relations. Caviedes and Cimino

[1] http://mbr.nlm.nih.gov/index.shtml.

[18] later evaluated this measure using the PAR/CHD relations. The *path* measure is a modification of this and is calculated as the reciprocal of the length of the shortest path.

Wu and Palmer [19] extend this measure by incorporating the depth of the Least Common Subsummer (LCS). The LCS is the most specific concept two concepts share as an ancestor. In this measure, the similarity is twice the depth of the two concepts LCS divided by the product of the depths of the individual concepts as defined in Eq. (1).

$$\mathrm{sim}_{wup} = \frac{2 * \mathrm{depth}(\mathrm{lcs}(c_1, c_2))}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)} \tag{1}$$

Leacock and Chodorow [20] extend the path measure by incorporating the depth of the taxonomy. Here, the similarity is the negative log of the shortest path (*minpath*) between two concepts divided by twice the total depth of the taxonomy (*D*) as defined in Eq. (2).

$$\mathrm{sim}_{lch} = -\log\frac{\mathrm{minpath}(c_1, c_2)}{2 * D} \tag{2}$$

Nguyen and Al-Mubaid [21] incorporate both the depth and LCS in their measure. In this measure, the similarity is the log of two plus the product of the shortest distance between the two concepts minus one and the depth of the taxonomy (*D*) minus the depth of the concepts' LCS (*d*) as defined in Eq. (3). Its range depends on the depth of the taxonomy.

$$\mathrm{sim}_{nam} = \log(2 + (\mathrm{minpath}(c_1, c_2) - 1) * (\mathrm{D} - \mathrm{d})) \tag{3}$$

3.1.2. IC-based

Information content (IC) is formally defined as the negative log of the probability of a concept. Resnik [22] modified IC to be used as a similarity measure. He defined the similarity of two concepts to be the IC of their LCS as shown in the following equation:

$$\mathrm{sim}_{res} = \mathrm{IC}(\mathrm{lcs}(c_1, c_2)) = -\log(P(\mathrm{lcs}(c_1, c_2))) \tag{4}$$

Jiang and Conrath [23] and Lin [24] extended Resnik's IC-based measure by incorporating the IC of the individual concepts. Lin defined the similarity between two concepts by taking the quotient between twice the IC of the concepts' LCS and the sum of the IC of the two concepts as shown in Eq. (5). This is similar to the measure proposed by Wu and Palmer; differing in the use of IC rather than the depth of the concepts.

$$\mathrm{sim}_{lin} = \frac{2 * \mathrm{IC}(\mathrm{lcs}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)} \tag{5}$$

Jiang and Conrath defined the distance between two concepts to be the sum of the IC of the two concepts minus twice the IC of the concepts' LCS. We modify this measure to return a similarity score by taking the reciprocal of the distance as shown in Eq. (6).

$$\mathrm{sim}_{jcn} = \frac{1}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 * \mathrm{IC}(\mathrm{lcs}(c_1, c_2))} \tag{6}$$

3.2. Information content

The information content of a concept can be calculated using information derived from a corpus (corpus-based) or information derived from a taxonomy (taxonomy-based). In this section, we describe both techniques.

3.2.1. Corpus-based

Information content is defined as the negative log of the probability of a concept. We define the probability of a concept, *c*, by summing the probability of the concept, *P*(*c*), occurring in some text plus the probability its descendants, *P*(*d*), occurring in the same text as seen in Eq. (7).

$$P(c*) = P(c) + \sum_{d \in descendant(c)} P(d) \tag{7}$$

The initial probability of a concept, *P*(*c*), and its descendants, *P*(*d*), is obtained by dividing the number of times a concept is seen in the corpus, *freq*(*d*), by the total number of concepts, *N*, as seen in Eq. (8).

$$P(d) = freq(d)/N \tag{8}$$

3.2.2. Taxonomy-based

The challenge with probability calculations for concepts is that a large number of annotations are required in order to provide sufficient coverage of the underlying taxonomy to achieve reasonable estimates. Intrinsic information content seeks to alleviate this problem while still capturing the generality and concreteness of a concept. It assess the informativeness of concept based on the its placement within the hierarchy by looking at its incoming (ancestors) and outgoing (descendant) links.

In this work, we use the Intrinsic IC calculation proposed by Sanchez et al. [25] defined in Eq. (9).

$$IC(c) = -log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{max\_leaves + 1}\right) \tag{9}$$

where *leaves* are the number of descendant of concept *c* that are leaf nodes, *subsumers* are the number of concept *c*'s ancestors and *max_leaves* are the total number of leaf nodes in the taxonomy.

3.3. Relatedness measures

Lesk [12] introduces a measure that determines the relatedness between two concepts by counting the number of overlaps between their two definitions. An overlap is the longest sequence of one or more consecutive words that occur in both definitions. When implementing this measure in WordNet, Banerjee and Pedersen [26] found that the definitions were short, and did not contain enough overlaps to distinguish between multiple concepts, therefore, they extended this measure by including the definition of the related concepts.

Patwardhan and Pedersen [27] extend this measure using second-order co-occurrence vectors. In this method, a vector is created for each word in the concepts definition containing words that co-occur with it in a corpus. These word vectors are averaged to create a single co-occurrence vector for the concept. The similarity between the concepts is calculated by taking the cosine between the concepts second-order vectors. Liu et al. [28] modify and extend this measure to be used to quantify the relatedness between biomedical and clinical terms in the UMLS.

4. Resources

In this section, we describe the resources used in our experiments.

4.1. Medline

Medline[2] is a bibliographic database containing over 18.5 million citations to journal articles in the biomedical domain and is maintained by National Library of Medicine. The 2009 Medline Baseline encompasses approximately 5200 journals starting from 1948 and contains 17,764,826 citations; consisting of 2,490,567 unique unigrams (single words) and 39,225,736 unique bigrams (two-word sequences). The majority of the publications are scholarly journals

[2] http://mbr.nlm.nih.gov/Download/index.shtml.

but a small number of newspapers and magazines are included.

4.2. Unified Medical Language System

The UMLS is a data warehouse containing three knowledge sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus contains approximately 2 million biomedical and clinical concepts from over 100 different terminologies that have been semi-automatically integrated into a single source. One such source is the Medical Subject Heading (MeSH) Thesaurus which is the National Library of Medicine's (NLM) controlled vocabulary thesaurus consisting of biomedical concepts created for the purposes of indexing and is used for indexing articles from the Medline database. The concepts in MeSH are organized in a hierarchical structure in order to permit searching at various levels of specificity. The concepts are connected by two main types of hierarchical relations: *parent/child* (PAR/CHD) and *broader/narrower* (RB/RN). The PAR/CHD relations are strictly *is-a* relations while the RB/RN relations contain *part* − *of* relations.

The Semantic Network consists of a set of broad subject categories called semantic types in which each concept in the Metathesaurus is assigned one or more semantic type. For example, the semantic type of C0206250 [Autonomic nerve] is *Body Part, Organ, or Organ Component*. Currently, there exist 135 semantic types in the Semantic Network.

The SPECIALIST Lexicon contains terms that are used in the biomedical and health-related domain along with linguistic information such as spelling variants. In this work, we use the SPECIALIST Lexicon to identify terms surrounding the ambiguous word in our dataset.

5. Method

UMLS::SenseRelate [29] is a freely available open source Perl package[3] developed to assign UMLS senses to ambiguous terms in biomedical text. UMLS::SenseRelate is an extension of WordNet::SenseRelate [1], a freely available software package[4] which is an implementation of Patwardhan et al.'s [1] method to disambiguate terms in general English.

In this method, each possible concept of a word is assigned a score by summing a weighted similarity score between it and the terms or concepts surrounding the ambiguous word in a given window of context. The concept with the highest score is assigned to the target word. We identify the terms surrounding the target word using the SPECIALIST Lexicon. The sequence of words with the longest match to the terms that exist in the lexicon are treated as a single term. Once the terms are identified, the algorithm computes the similarity or relatedness between the possible concept of the target word and each of the surrounding terms using the freely available open source Perl package UMLS::Similarity[5] developed to calculate the similarity or relatedness between biomedical terms. If a surrounding term is polysemous, the algorithm uses the concept that returns the highest similarity score. The score is then weighted based on how far it is from the target word by multiplying the reciprocal of its distance to the similarity score.

To provide an example, consider the following sentence containing the target word *tolerance* which has the possible concepts Drug Tolerance [C0013220] and an Immune Tolerance [C0020963]: It attenuates *tolerance* to analgesic effect of morphine in mice with skin cancer.

In this example, we use a window size of five which refers to five content terms to the right and the left of the target word and attempt to map them to CUIs. In this case, the content words are: *attenuates*, *analgesic*, *effect*, *morphine*, *mice*, *skin cancer*. Of these six words, only three have mappings to CUIs in MeSH: *morphine*:C0026549, *mice*:C0026809, and *skin cancer*:C0007114. In this method, we treat *skin cancer* as a single term mapping to the concept C0007114 rather than individual words which would map to *skin*:C1123023 and *cancer*:C0006826.

The WSD algorithm then obtains similarity scores between each of the possible concepts and the concepts of the content words in the window of context. Each score is then multiplied by the reciprocal of its distance from the target words, for example, *skin cancer* is five content words away from *tolerance* and therefore multiplied by $\frac{1}{5}$. The scores are then summed to obtain a total score for each possible concept as shown in Fig. 1. The pseudocode for this algorithm is in Algorithm 1.

Algorithm 1. SenseRelate algorithm

```
procedure SenseRelate($concepts$, $window$, $distance$)
  $score \leftarrow 0$
  $annotation \leftarrow null$
  for each concept in concepts do
    $sum \leftarrow -1$
    $numberterms \leftarrow 0$
    for $i = 0 \rightarrow window.length$ do
      $term \leftarrow window[i]$
      $cuis \leftarrow getUMLSConcepts(term)$
      $maxscore \leftarrow -1$
      for each cui in cuis do
        $similarity \leftarrow getRelatedness(cui, concept)$
        If $similarity > maxscore$ then
          $maxscore \leftarrow similarity$
        end if
      end for
      If $maxscore > -1$ then
        $sum \leftarrow sum + \left(maxscore * \frac{1}{distance[i]}\right)$
        $numberterms \leftarrow numberterms + 1$
      end if
    end for
    $sum \leftarrow \frac{sum}{numberterms}$
    If $sum > score$ then
      $score \leftarrow sum$
      $annotation \leftarrow concept$
    end if
  end for
  return $annotation$
end procedure
```

6. Data

6.1. Evaluation data

We evaluate our method on NLM's *MSH-WSD* dataset developed by Jimeno-Yepes et al. [13]. The data set contains 203 ambiguous terms and acronyms from the 2010 Medline baseline. Each instance of a term was automatically assigned a CUI from the 2009AB version of the UMLS by exploiting the fact that each instance in Medline is manually indexed with Medical Subject Headings in which each heading has an associated CUI. Each target word contains approximately 187 instances, has 2.08 possible concepts and has a 54.5% majority sense. Out of 203 target words, 106

[3] http://search.cpan.org/dist/UMLS-SenseRelate/.

[4] http://search.cpan.org/dist/WordNet-SenseRelate/.

[5] http://search.cpan.org/dist/UMLS-Similarity/.

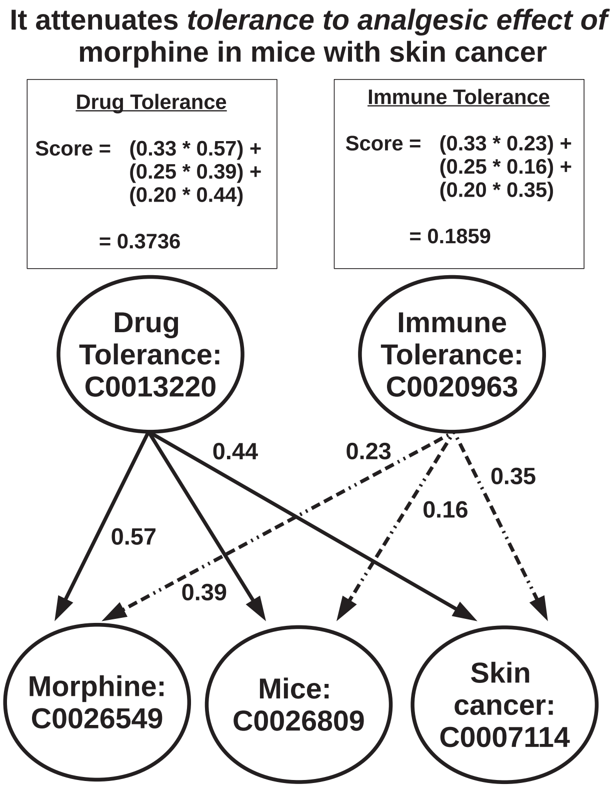
It attenuates *tolerance to analgesic effect of* morphine in mice with skin cancer
Drug Tolerance
Score = (0.33 * 0.57) + (0.25 * 0.39) + (0.20 * 0.44)
= 0.3736
Immune Tolerance
Score = (0.33 * 0.23) + (0.25 * 0.16) + (0.20 * 0.35)
= 0.1859
Drug Tolerance: C0013220
Immune Tolerance: C0020963
0.44
0.23
0.35
0.16
0.57
0.39
Morphine: C0026549
Mice: C0026809
Skin cancer: C0007114

Fig. 1. Example of UMLS::SenseRelate method.

are terms, 88 are acronyms, and 9 have possible concepts that are both acronyms and terms. For example, the target word *cold* has the acronym *Chronic Obstructive Airway Disease* as a possible concept, as well as the term *Cold Temperature*. The total number of instances is 37,888.

6.2. IC similarity measure data

The IC similarity measure data is used to calculate the probability of a concept occurring in a corpus. We use the *UMLSonMedline* dataset created by NLM which consists of concepts from the 2009AB UMLS and the number of times they occurred in a snapshot of Medline taken on 12 January, 2009. The frequency counts were obtained by using the Essie Search Engine proposed by Ide et al. [30] which queried Medline with normalized strings from the 2009AB MRCONSO table in the UMLS. The frequency of a CUI was obtained by aggregating the frequency counts of the terms associated with the CUI to provide a rough estimate of its frequency. The IC measures use this information to calculate the probability of a concept.

6.3. Relatedness measure data

The relatedness measure data is used by the *vector* measure to build the second-order co-occurrence matrix. We evaluate building the matrix using three different data sources: NLM Medline Bigram Data, UMLS MRCOC Co-occurrence Data and MSH-WSD Medline Data.

6.3.1. NLM Medline Bigram Data

The NLM Medline Bigram data[6] consists of bigram counts obtained from the 2012 Medline baseline. The bigrams are collected using a sliding window method where a bigram consists of a *valid* word and current word to create the bigram. A valid word is any word that does not consists of all numbers, does not contain non-ascii characters and is not a stop word. At a line break, the process starts over therefore a bigram will not contain words across two citations. The bigrams are collected from the title and abstract fields in the medline citation. There are 39,225,736 bigrams in the 2012 Medline baseline.

6.3.2. UMLS MRCOC Co-occurrence Data

The UMLS Co-occurrence table (MRCOC) is part of the UMLS Metathesaurus. It contains information about the co-occurrence of concepts "that were designated as principal or main points in the same journal article" [31]. In the 2012AA version of the UMLS, MRCOC contains 20,779,850 CUI pairs including reciprocal information.

6.3.3. MSH-WSD Medline Data

The MSH-WSD is a subset of the 2009 Medline baseline. The bigrams are extracted from the dataset using Text::NSP[7] where stopwords and non-alpha character words are removed. The newline function is also used so a bigram will not contain words across two different abstracts. There are 1,312,413 bigrams in the MSH-WSD Medline data.

7. Experimental framework

In this article, we evaluate each of the similarity and relatedness measures previously discussed on the task of WSD using UMLS::SenseRelate. We also evaluate the following parameters:

- **Measures**: we compare the similarity and relatedness measures discussed in Section 3
- **Window size**: we explore various window sizes in which the terms surrounding the ambiguous word are obtained
- **Weighting**: we explore weighting the similarity and relatedness scores based on their distance from the target word
- **Vector data**: we explore obtaining vector data from: NLM Medline Bigram Data, UMLS MRCOC Co-occurrence Data, and MSH-WSD Medline Data

These experiments were conducted using the 2012AB version of the UMLS. We use the MeSH taxonomy located in the UMLS Metathesaurus for the similarity measures and the entire UMLS (Level 1 + SNOMED CT) for the relatedness measures. Differences between the means of disambiguation accuracy produced by various approaches were tested for statistical significance using pair-wise Student's t-test. The programs to evaluate the results and calculate the significance are included in the UMLS::SenseRelate package.

With respect to the vector data, the UMLS MRCOC contains 20,779,850 CUI pairs with their frequency counts. We extracted the CUIs Preferred Term and utilized that in place of its CUI. We limited those bigrams by including those that occur more than 50 times but less than 1000. This reduced the total number of bigrams to 15,786,429 pairs. The NLM Medline Bigrams are obtained from the entirety of Medline and consists of approximately 40 million unique bigrams. We limited those bigrams by including those in which both words in the bigram map to a CUI in the UMLS and occur more than 50 times but less than 1000. This reduced the total number of bigrams to 45,692. The MSH-WSD Medline Data bigrams are obtained from the MSH WSD data. We limited those bigrams by including only those that occurred more than 50 times but less than 1000 (as with the above datasets). This reduced the total number of bigrams to 2142. We also explored various lower

[6] http://mbr.nlm.nih.gov/Download/index.shtml.

[7] http://search.cpan.org/dist/Text-NSP/.

Table 1
Accuracy of measures across Window Sizes (WS).

WS	Path-based				Corpus-based IC			Taxonomy-based IC			Relatedness	
	path	lch	wup	nam	res	jcn	lin	res	jcn	lin	a-lesk	vec
2	0.63	0.63	0.64	0.63	0.64	0.65	0.65	0.65	0.65	0.64	0.67	0.68
5	0.66	0.66	0.67	0.66	0.68	0.69	0.69	0.68	0.68	0.68	0.68	0.68
10	0.68	0.68	0.69	0.69	0.70	0.71	0.71	0.70	0.70	0.70	0.68	0.67
25	0.71	0.69	0.71	0.71	0.73	0.73	0.74	0.73	0.73	0.73	0.66	0.65
50	0.71	0.69	0.70	0.71	0.73	0.74	0.74	0.73	0.73	0.73	0.66	0.65
70	0.71	0.69	0.70	0.72	0.73	0.74	0.74	0.73	0.73	0.73	0.66	0.65

Table 2
Significance of measures using window size of 25.

		Path-based				Corpus-based IC			Taxonomy-based IC			Relatedness	
		path	lch	wup	nam	res	jcn	lin	res	jcn	lin	a-lesk	vec
Baseline	MSB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Path-based	path		0.35	0.99	0.78	0.12	0.06	0.03	0.03	0.08	0.03	0.0008	0.00
	lch			0.45	0.25	0.02	0.007	0.002	0.01	0.005	0.005	0.01	0.00
	wup				0.69	0.11	0.05	0.02	0.16	0.09	0.11	0.001	0.00
	nam					0.23	0.11	0.05	0.14	0.08	0.08	0.0002	0.00
Corpus-based IC	res						0.71	0.51	0.77	0.79	0.94	0.00	0.00
	jcn							0.75	0.52	0.94	0.67	0.00	0.00
	lin								0.34	0.71	0.46	0.00	0.00
Taxonony based IC	res									0.57	0.82	0.00	0.00
	jcn										0.73	0.00	0.00
	lin											0.00	0.00
Relatedness	a-lesk												0.23

and upper bound cutoff combinations with lower bound cutoffs of 2, 10, 25, 50, 100, 500 and 1000, and upper bound cutoffs of 1000 and the entire set. We found there was no difference in accuracy when using a lower bound cutoff of 50 or 100 for each of the datasets, and using a lower bound cutoff of 2, 10 or 25 resulted in lower accuracies.

8. Results and discussion

Table 1 shows the accuracy of UMLS::SenseRelate using: the path measure (path), the path-based measures proposed by Leacock and Chodorow (lch), Wu and Palmer (wup) and Nguyen and Al-Mubaid (nam); the IC-based measures proposed by Resnik (res), Jiang and Conrath (jcn) and Lin (lin) using both the corpus-based and taxonomy-based information content; the adapted lesk (a-lesk) relatedness measure proposed by Banerjee and Pedersen [26] (a-lesk); and the relatedness measure proposed by Patwardhan and Pedersen [27] (vec). The results are shown when using window sizes of 2, 5, 10, 25, 50 and 70. For reference, a window size of two indicates the algorithm is using two content terms the left and two content terms to the right of the target word to determine the appropriate concept.

The overall results show that IC measures using either the taxonomy or corpus-based IC consistently obtain a higher disambiguation accuracy than the other measures. Table 2 shows the statistical significance between each of the measures and the majority sense baseline (MSB). This baseline is often used to evaluate supervised learning algorithms and indicates the accuracy that would be achieved by assigning the most frequent concept to every instance. The majority sense baseline for the MSH-WSD dataset is 0.5448. The results show that for each measure, the disambiguation accuracy is statistically significantly greater than the baseline.

The results using *lin* with the corpus-based IC show that the measure obtains a statistically significantly higher disambiguation accuracy than path-based and relatedness measures. There is not a statistical significance between *lin* and the other IC-based measures. This is also the case for *lin* when using the taxonomy-based IC except when compared to the nam where the statistical significance is just under 95%.

The overall IC results show that for *res*, *jcn* and *lin* the disambiguation accuracy is the either the same or not statistically significant regardless when using the corpus-based or taxonomy-based IC.

The windowing results show that for the similarity measures the accuracy plateaus after a window size of 25. This observation is similar to those identified by McInnes et al. [2] when using corpus-based IC and Garla and Brandt [3] when using Taxonomy-based IC. The windowing results for the relatedness measures show that the disambiguation accuracy plateaus after a window size of 5. Analysis of these results show that the similarity and relatedness cannot be computed for some CUIs in the window, and this varies depending on the measure and the window size. Table 3 show the average number of relatedness or similarity scores obtained for each type of measure over the window sizes. For example, when using the path-based measures and a window size of 25, we could only compute the similarity of 8.12 of the content terms in the window.

Table 3
Mappings of surrounding words to CUIs.

WS	Path-based	Corpus-based IC	Taxonomy-based IC	Relatedness
2	0.82	0.79	0.83	2.37
5	1.97	1.85	1.97	6.28
10	3.71	3.47	3.72	12.34
25	8.12	7.57	8.13	28.11
50	13.61	12.63	13.61	45.47
70	16.78	15.56	16.78	63.14

Table 4
Accuracy measures using weighting.

WS	Path-based				Corpus-based IC			Taxonomy-based IC			Relatedness	
	path	lch	wup	nam	res	jcn	lin	res	jcn	lin	a-lesk	vec
2	0.63	0.63	0.64	0.63	0.64	0.65	0.65	0.64	0.64	0.64	0.67	0.68
5	0.66	0.66	0.67	0.66	0.68	0.69	0.69	0.68	0.68	0.68	0.69	0.69
10	0.69	0.68	0.69	0.69	0.71	0.71	0.72	0.70	0.71	0.71	0.69	0.70
25	0.71	0.70	0.71	0.71	0.73	0.74	0.74	0.73	0.74	0.73	0.69	0.70
50	0.72	0.70	0.71	0.71	0.74	0.74	0.75	0.74	0.75	0.74	0.69	0.70
70	0.73	0.70	0.72	0.73	0.75	0.75	0.75	0.75	0.75	0.75	0.69	0.70

Table 5
Accuracy on MSH-WSD using vector with different data sources.

Dataset	Window size					
	2	5	10	25	50	70
NLM Medline Bigrams	0.68	0.69	0.69	0.69	0.69	0.69
UMLS MRCOC	0.68	0.67	0.64	0.63	0.63	0.63
MSH-WSD Medline Data	0.68	0.69	0.70	0.70	0.70	0.70

Table 7
Comparison with Jimeno-Yepes et al. [13] on MSH-WSD dataset.

Dataset	Jimeno-Yepes et al.		UMLS
	MRD	2-MRD	SenseRelate
Abbreviation set	0.87	0.85	0.83
Term set	0.72	0.68	0.67
Term/abbreviation set	0.88	0.94	0.77
Overall MSH-WSD set	0.81	0.78	0.75

These results indicate that a small number of locally occurring terms provide a sufficient enough of a distinction to determine of the concept of the target word. This is consistent with the finding reported by Choueka and Lusignan [32] who found that only a small window size was needed for humans to determine the appropriate concept of an ambiguous word.

The results show that the number of scores used in the overall calculation is slightly higher for the path-based and taxonomy-based IC measures than the corpus-based IC measures. This is because not all concepts in MeSH were seen in our corpus and the information content could not be calculated. For example, the concept for *drug induced liver injury* (C2717837) was not found in our corpus and has an information content of zero.

The results also show that the number of scores used in the overall calculation is higher when using the relatedness measures over the similarity measures. This is because the relatedness measures utilize definitional information obtained from the entire UMLS where as similarity measures are using the path information only from MeSH. Vector methods in general are subject to noise introduced by features that are not able to distinguish between the different concepts of a target word. These results indicate that as the number of mappings increases the amount of noise increases degrading the disambiguation accuracy of the algorithm.

To address this, we incorporate a weighting mechanism that weights the surrounding term based on its distance from the target word. Table 4 shows the results using weighting. A comparison between Tables 4 and 1 show that the disambiguation accuracy of the similarity results for the path-based measures increased slightly although the increase is not significant ($p \leqslant 0.01$); the IC measures showed no change in accuracy; and lastly, the relatedness measures showed a statistically significant increase in disambiguation accuracy ($p \leqslant 0.02$).

As discussed above, the relatedness measures contain a larger number of CUI pairs whose relatedness can be quantified. We believe that this introduces noise degrading the accuracy of the results. Weighting score based on the distance concept is from the target word reduces the amount of noise in the relatedness measures, increasing the accuracy of the results.

The vector measure uses bigram or co-occurrence information obtained from a corpus. Table 5 shows the accuracy of the vector measure evaluated on the three different corpora described in Section 7: (1) Medline Bigram Data provided by NLM; (2) UMLS MRCOC and (3) MSH-WSD Medline Bigrams. The table also shows the accuracy when using a window size of 2, 5, 10, 25, 50 and 70; weighting was used for these results.

The results show that using the MSH-WSD Medline data obtains the highest accuracy although these results are not statistically significant when compared to the NLM Medline Bigrams ($p \leqslant 0.01$). The bigram information from MSH-WSD is also contained in the NLM Medline Bigrams. We believe that using the bigram information from MSH-WSD obtained a slightly higher results though because the Medline bigrams encompass a larger amount of information over a longer period of time. The information from UMLS MRCOC obtained the lowest results. The terms from the UMLS MRCOC table consist of terms that co-existed together in a journal article. We believe that the context (window) in which the terms are identified is too large for sense discrimination.

9. Comparison with previous work

In this section, we compare our results with that of previous work. Table 6 shows the results obtained by Garla and Brandt [3] over a subset of the similarity measures described in Section 3.1.

The results from Table 6 show that all of the measures except for *jcn* accuracy reported by Garla and Brandt [3] are approximately close to those obtained using UMLS::SenseRelate. We believe the differences in results are due to two factors: (1) different versions of the UMLS and (2) the identification of the concepts of the surrounding terms. The authors evaluate obtaining concepts of the surrounding terms using the Named Entity Recognition

Table 6
Comparison with Garla and Brandt [3] on MSH-WSD dataset.

	Path-based				Corpus-based IC			Taxonomy-based IC		
	path	lch	wup	nam	res	lin	jcn	res	jcn	lin
Garla and Brandt	0.77	0.70	0.72						0.81	0.76
UMLS::SenseRelate	0.73	0.70	0.72	0.73	0.75	0.75	0.75	0.75	0.75	0.75

(NER) method from the clinical Text Analysis and Knowledge extraction System (cTAKES) and MetaMap. Their results for the MSH-WSD data set indicated that using MetaMap obtained a higher disambiguation accuracy.

We also compare UMLS::SenseRelate using lin with a window size of 25 with those reported by Jimeno-Yepes et al. [13]. As discussed in Section 6.1, the MSH-WSD dataset can be broken up into (1) terms (Term Set), (2) acronyms (Abbreviation Set), and (3) concepts that are both acronyms and terms (Term/Abbreviation Set). Table 7 show the results of UMLS::Similarity on the breakdown of the MSH-WSD dataset with those results reported by Jimeno-Yepes et al. [13].

The results show that UMLS::SenseRelate obtains accuracies comparable to that of MRD and 2-MRD on the Abbreviation Set; has a lower accuracy than MRD on the Term Set and Overall MSH-WSD Set; and is comparable with 2-MRD on the Term Set and Overall MSH-WSD Set.

As described in Section 4, the MRD and 2-MRD methods consist of creating a vector representing the target word based on the surrounding content words, and a vector representing the concept based on its definition. The cosine is calculated between the target word vector and each possible concept's vector. The concept vector closest to the target word vector is assigned to the target word. The limitation to this method is a definition is required for each possible concept and not all concepts have a definition in the UMLS. The authors attempt to alleviate this problem by using the definitions of related concepts but this dilutes the actual meaning of the concept and creates the possibility of two non-synonymous concept having the same definition. This limitation does not exist using SenseRelate algorithm with the similarity measures.

10. Conclusions and future work

In this article, we evaluated a knowledge-based method for WSD, called UMLS::SenseRelate. The benefit of this method is that it does not require manual annotation and yields a disambiguation accuracy sufficiently high for most practical purposes. The objective of this work was to evaluate a method that can disambiguate terms in biomedical text using similarity and relatedness information extrapolated from the UMLS, and evaluate the efficacy of similarity and relatedness measures. To do this, we evaluated UMLS::SenseRelate on the MSH-WSD dataset using semantic similarity and relatedness measures from the UMLS::Similarity package. We found that on this dataset IC-based measures obtain a statistically significantly higher overall disambiguation accuracy than path-based measures and relatedness measures. We believe this is because the IC-based measures weight the path based on where it exists in the taxonomy using the probability of the concepts occurring in a corpus of text (corpus-based), or is location with respect to the other concepts in the taxonomy (taxonomy-based).

The results obtained by Garla and Brandt [3] also showed that the IC measure obtained a significantly higher disambiguation accuracy than the path based measures on the MSH-WSD dataset. The authors did not find this to be the case when evaluating it on the NLM-WSD dataset [33]. In the future, we would like to evaluate this method on additional datasets such as NLM-WSD and the Abbrev dataset [34].

Currently, the terms are obtained from the SPECIALIST Lexicon and are mapped to concepts using a dictionary look up, In the future, we plan to incorporate a more comprehensive method that incorporates phrasal information.

In this method, the selection of the most appropriate sense is determined by a simple averaging of the weighted similarity scores of the surrounding terms. In the future, we plan to look at other aggregation criteria such as order weighted averaging operators which have been previously used in decision making schemes for aggregating uncertain information with uncertain weights.

11. Data

The MSH-WSD dataset described in Section 6.1 can be obtained from http://wsd.nlm.nih.gov.

Acknowledgments

Portions of this work were supported by the National Institute of Health, National Library of Medicine Grant #R01LM009623-01.

We would like to the National Library of Medicine for providing the MSH-WSD dataset.

References

[1] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics; 2003. p. 241–57.

[2] McInnes B, Pedersen T, Liu Y, Pakhomov S, Melton G. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In: Proceedings of the American medical informatics association symposium, Washington, DC; 2011.

[3] Garla VN, Brandt C. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. J Am Med Inform Assoc 2012;0(5):1–5.

[4] Zhong Z, Ng H. It makes sense: a wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 system demonstrations, association for computational linguistics; 2010. p. 78–83.

[5] Stevenson M, Guo Y, Gaizauskas R, Martinez D. Disambiguation of biomedical text using diverse sources of information. BMC Bioinformatics 2008;9(Suppl. 11):11.

[6] Brody S, Lapata M. Bayesian word sense induction. In: Proceedings of the 12th conference of the European chapter of the association for computational linguistics; 2009. p. 103–11.

[7] Pedersen T. The effect of different context representations on word sense discrimination in biomedical texts. In: Proceedings of the 1st ACM international health informatics symposium; 2010. p. 56–65.

[8] Navigli R, Faralli S, Soroa A, de Lacalle O, Agirre E. Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM; 2011. p. 2317–20.

[9] Humphrey S, Rogers W, Kilicoglu H, Demner-Fushman D, Rindflesch T. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. J Am Soc Inform Sci Technol 2006;57(1):96–113. doi:http://dx.doi.org/10.1002/asi.v57:1.

[10] Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, et al. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics 2009;10(1):28.

[11] Jimeno-Yepes A, Aronson A. Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC Bioinformatics 2010;11(1):569.

[12] Lesk M. 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; 1986. p. 24–6.

[13] Jimeno-Yepes A, McInnes B, Aronson A. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. BMC Bioinform 2011;12(1):223.

[14] McInnes BT. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: student research workshop. Association for Computational Linguistics; 2008. p. 49–54.

[15] Agirre E, Soroa A, Stevenson M. Graph-based word sense disambiguation of biomedical documents. Bioinformatics 2010;26(22):2889–96.

[16] Stevenson M, Agirre E, Soroa A. Exploiting domain information for word sense disambiguation of medical documents. J Am Med Inform Assoc 2012;19(2):235–40.

[17] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;19(1):17–30.

[18] Caviedes J, Cimino J. Towards the development of a conceptual distance metric for the umls. J Biomed Inform 2004;37(2):77–85.

[19] Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proceedings of the 32nd meeting of association of computational linguistics; 1994. p. 133–8.

[20] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. WordNet: An Electron Lexical Database 1998;49(2):265–83.

[21] Nguyen H, Al-Mubaid H. New ontology-based semantic similarity measure for the biomedical domain. In: Proceedings of the IEEE international conference on granular computing; 2006. p. 623–8.

[22] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence; 1995. p. 448–53.

[23] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings on international conference on research in computational linguistics; 1997. p. 19–33.

[24] Lin D. An information-theoretic definition of similarity. In: Proceedings of the international conference on machine learning; 1998. p. 296–304. <http://citeseer.ist.psu.edu/95071.html>.

[25] Sánchez D, Batet M, Isern D. Ontology-based information content computation. Knowledge-Based Syst 2011;24(2):297–303.

[26] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th international joint conference on AI; 2003. p. 805–10.

[27] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 workshop making sense of sense – bringing computational linguistics and psycholinguistics together, Trento, Italy; 2006. p. 1–8.

[28] Liu Y, McInnes B, Pedersen T, Melton-Meaux G, Pakhomov S. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, umls and wordnet. In: Proceedings of the 2nd ACM SIGHIT symposium on international health informatics. ACM; 2012. p. 363–72.

[29] McInnes B, Pedersen T, Pakhomov S. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: Proceedings of the American medical informatics association symposium, San Fransico, CA; 2009.

[30] Ide N, Loane R, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.

[31] Zeng Q, Cimino J. Automated knowledge extraction from the umls. In: Proc AMIA Symp. American Medical Informatics Association; 1998. p. 568.

[32] Choueka Y, Lusignan S. Disambiguation by short contexts. Comput Humanit 1985;19(3):147–57.

[33] Weeber M, Mork J, Aronson A. Developing a test collection for biomedical word sense disambiguation. In: Proceedings of the American medical informatics association symposium, Washington, DC; 2001. p. 746–50.

[34] Stevenson M, Guo Y, Al Amri A, Gaizauskas R. Disambiguation of biomedical abbreviations. In: Proceedings of the ACL BioNLP workshop; 2009. p. 71–9.