Exploiting domain information for Word Sense Disambiguation of medical documents

Mark Stevenson,[1] Eneko Agirre,[2] Aitor Soroa[2]

[1]Department of Computer Science, Sheffield University, Sheffield, UK
[2]IXA NLP Group, University of the Basque Country, Donostia, Basque Country, Spain

Correspondence to
Dr Mark Stevenson, Department of Computer Science, Sheffield University, Regent Court, 211 Portobello, Sheffield S1 4DP, UK;
m.stevenson@dcs.shef.ac.uk

Received 1 June 2011
Accepted 11 August 2011
Published Online First
7 September 2011

ABSTRACT

Objective Current techniques for knowledge-based Word Sense Disambiguation (WSD) of ambiguous biomedical terms rely on relations in the Unified Medical Language System Metathesaurus but do not take into account the domain of the target documents. The authors' goal is to improve these methods by using information about the topic of the document in which the ambiguous term appears.

Design The authors proposed and implemented several methods to extract lists of key terms associated with Medical Subject Heading terms. These key terms are used to represent the document topic in a knowledge-based WSD system. They are applied both alone and in combination with local context.

Measurements A standard measure of accuracy was calculated over the set of target words in the widely used National Library of Medicine WSD dataset.

Results and discussion The authors report a significant improvement when combining those key terms with local context, showing that domain information improves the results of a WSD system based on the Unified Medical Language System Metathesaurus alone. The best results were obtained using key terms obtained by relevance feedback and weighted by inverse document frequency.

INTRODUCTION

The published literature in medicine and related fields now forms a vast amount of information which is so large that it can only be accessed effectively using automatic search tools.[1 2] Providing these tools is an important goal since access to information in the medical literature has been shown to be beneficial for both consumers and health professionals.[3 4] Automatic processing of biomedical documents is, however, made difficult by the fact that they contain terms that are ambiguous. For example, 'culture' can mean 'laboratory procedure' (eg, 'In peripheral blood mononuclear cell culture') or 'anthropological culture' (eg, 'main accomplishments of introducing a quality management culture').

The process of resolving lexical ambiguities is known as Word Sense Disambiguation (WSD) and has been widely studied in Natural Language Processing.[5 6] Several approaches to WSD in the biomedical domain have been based on supervised methods.[7–9] However, these rely on large datasets for training which are difficult to obtain or create.[10] Recently, researchers have explored techniques for automatically identifying examples and using them as an alternative to manually labeled data,[11 12] although these approaches have yet to be applied to more than small sets of ambiguous terms. Humphreys *et al*[13] avoided the need for labeled data by making use of Journal Descriptors[14] to exploit information about the topic of the document in which an ambiguous term appears. This approach assigned ambiguous terms one of the 135 Semantic Types from the Unified Medical Language System (UMLS) Metathesaurus[15] but was unable to distinguish between meanings that have the same Semantic Type.

Unlike supervised approaches, knowledge-based approaches do not require training data and make use of information from some external resource, or knowledge base (KB). McInnes[16] reported an approach that used the UMLS Metathesaurus as a KB and could distinguish between all possible meanings (and not just those with different Semantic Types). Textual descriptions for each meaning of an ambiguous word were generated from the Metathesaurus and the most appropriate sense chosen by identifying the one which shared the most terms with the context, a commonly used technique for WSD.[17] As an alternative, graph-based techniques have recently proved to be a successful knowledge-based approach.[18 19] These have recently been applied to the biomedical domain by creating a graph using the relations in the UMLS Metathesarurus as a KB and then applying a random-walk algorithm in order to determine the most appropriate meaning according to the context.[11 18] Jimeno-Yepes and Aronson[11] compared a number of knowledge-based WSD algorithms and found that the graph-based approach is outperformed by alternative approaches. However, a similar approach[20] reported superior performance using a more recent version of UMLS.

Most work on WSD has ignored the domain of the target documents. More recently, there has been interest in methods that take into account the domain in which an ambiguous word appears.[21–23] Medical Subject Heading (MeSH) terms are manually curated labels for biomedical and health-related documents that often provide information about the topic of the document to which they are applied. Several studies have shown that MeSH terms are useful for WSD of biomedical documents.[9 24–28] However, all of these approaches have used MeSH terms as features in a supervised (or semisupervised) system. This paper makes use of MeSH terms within a knowledge-based approach by using them to create a set of key terms closely associated with each MeSH term. These key terms are used as context for a WSD system, and it is found that combining these key terms with local context outperforms the use of either in isolation,

and significantly improves the system's performance. The main advantage of our system with respect to previous WSD systems for biomedical documents, for example,[13 16 20] is that it makes use of information contained in the UMLS Metathesaurus while combining it with information about the domain of the target document automatically learned from text.

BACKGROUND

This section describes the graph-based methods for WSD as used in Agirre *et al*.[20]

Graph-based WSD

A KB is typically formed by a set of concepts, relations among the concepts and a dictionary (a list of words linked to at least one concept). We consider the KB as a graph G=(V;E), using vertices V for representing concepts, and edges E for relations between them.

The WSD system is based on random walks over a graph representing a KB and uses the PageRank[29] algorithm, originally developed to identify important pages in web searches. PageRank can be viewed as a technique for scoring the vertices V according to their importance in the overall structure of the graph. After the PageRank calculation, the final weight assigned to a node represents the proportion of time that a random walker spends visiting that node after a sufficiently long time.

Assume that G has N vertices ($v_1,\ldots,v_n$). For a given vertex v_i, let $In(v_i)$ be the set of vertices pointing to it, and let d_j be the out-degree of vertex v_j. The PageRank of vertex v_i is defined as:

$$P(v_i) = c \sum_{v_j \in In(v_i)} \frac{1}{d_j} P(v_j) + (1-c)\frac{1}{N} \quad (1)$$

where c is the so-called damping factor, a scalar value between 0 and 1.

In standard PageRank, weight is assigned to unconnected vertices with probability 1−c and uniformly distributed across the graph, whereas for Personalized PageRank (PPR) it is chosen non-uniformly and specified by a teleport vector.[30] In order to introduce PPR, equation 1 is rewritten in a compact form using matrices. Let M be an N×N transition probability matrix, where $M_{ji}=1/d_i$ if a link from v_i to v_j exists, and zero otherwise. Let **v** be a stochastic normalized N×1 vector whose elements are all 1/N (the teleport vector). Then, the calculation of the PageRank Vector **P** over the graph G is equivalent to resolving the following equation:

$$P = cMP + (1-c)v \quad (2)$$

PPR is used for WSD by constructing a vector v that assigns high probabilities to the context words that surround the ambiguous word. Let $W=\{W_1,\ldots, W_m\}$ be an input context comprising words which have an entry in the dictionary, and can therefore be related to KB concepts. For each target word W_i, the system concentrates the teleport vector in the concepts of the words surrounding W_i, but not in the concepts of the target word itself, and applies the PPR over the graph. The target word is then disambiguated by choosing the concept associated with it with the highest rank. This approach has been used successfully by a number of authors.[19 20 31]

Application to UMLS

PPR has been applied to the disambiguation of medical documents using the UMLS Metathesaurus as a KB.[11 20] The UMLS was created by unifying a diverse range of controlled vocabularies and classification systems. It consists of more than one million biomedical concepts and five million concept names. The Metathesaurus is organized around concepts, and each is assigned a Concept Unique Identifier (CUI). Strings are considered ambiguous in the UMLS if they are associated with more than one CUI. For example, the following CUIs are associated with the term 'culture': C0010453 'Anthropological Culture' (eg, 'a quality management culture') and C0430400 'Laboratory culture' (eg, 'blood mononuclear cell culture'). The Metathesaurus also contains information about a wide variety of relations between CUIs in database tables. For example, the MRREL table relates C0010453 to C0015032 'Ethnology' and C0037455 'Societies'. A graph is created using the CUIs as vertices and edges between them defined using the MRREL table.

The dictionary contains mappings from words and phrases in text to UMLS CUIs. It is created using the MetaMap program[32] which splits the input text into phrases and maps each onto the set of possible CUIs that they could refer to, known as candidates. The set of candidates for each word or phrase in the context of the ambiguous terms is extracted from the MetaMap output and used to create the dictionary to define the possible CUIs for each word in its context.

The graph and dictionary were constructed using publicly available software (http://ixa2.si.ehu.es/ukb/) and resources (https://uts.nlm.nih.gov/) and can be easily replicated, as explained in Agirre *et al*.[20]

National Library of Medicine-WSD data set

The National Library of Medicine (NLM)-WSD data set[33] was used for evaluation. This is a collection of 50 terms that are ambiguous in the UMLS Metathesaurus and occur frequently in Medline. A hundred instances of each of the 50 terms were selected from citations added to Medline in 1998 and manually disambiguated. In addition to the meanings defined in UMLS, annotators had the option of assigning a special tag ('none') when none of the meanings in UMLS were judged to be appropriate. Following common practice among researchers who use this corpus,[11 13 16 20] we removed these instances from the test set leaving 3983 instances and 49 ambiguous terms. One term, association, was excluded, since all instances were labeled 'none.'

EXTRACTING DOMAIN TERMS

This section describes two general methods for identifying domain key terms; we then describe how key terms are created for the NLM-WSD dataset (Section 'Identifying key terms for NLM-WSD') as well as providing examples.

Domain context

Our approach to integrating domain information is to identify terms that are indicative of the domain (key terms) and use them as context, either as a replacement for or in addition to the local context. For example, the fact that 'culture' occurs in documents discussing microbiology is a strong indicator that it means 'cell culture' rather than one of the alternative meanings. Examples of key terms in the microbiology domain could include 'cell,' 'activity,' 'inhibited,' and 'assay.' We use these terms as additional context in the PPR algorithm.

This approach relies on being able to identify the key terms that indicate a domain. We use two different lexical statistics, which rely on the assumption that we have access to a corpus in which documents have domain labels associated with them. The domain labels we use are MeSH terms.[15] MeSH is a controlled vocabulary for indexing biomedical and health-related information. The most recent version contains 26 142 descriptors. MeSH

terms are manually assigned to Medline abstracts by human indexers and provide an accurate information about the domain and topic of the abstract. For example, MeSH terms associated with the abstract containing the phrase 'blood mononuclear cell culture' include Cultured Cells, Membrane Proteins, and Human.

Log likelihood

The first method for identifying key terms is the log-likelihood ratio, G^2, which has been widely used in language processing.[34–36] This approach relies on analyzing variables and assigning high scores when their co-occurrence is greater than would be expected by chance. The log-likelihood ratio has been used within a corpus comparison method to identify the terms that were indicative of each corpus,[37] and we reapply their method here. In this application, the variables are the occurrence (or otherwise) of a term within a document and assignment (or otherwise) of a domain label to a document. Terms are assigned high scores in relation to a domain label when the probability of their occurring in documents marked with the domain label is greater than chance would predict.

The G^2 score for each term and MeSH code is computed by creating a 2×2 contingency table listing the observed occurrences of that term in documents labeled with that MeSH code. An example table is shown in table 1, where o_{++} indicates the number of times a term occurs in a document labeled with the MeSH code, o_{+-} the number of times that it occurs in a document not labeled with the code, and so on.

The expected value for each cell in the contingency table can then be computed using equation 3, which allows the G^2 statistic to be calculated according to equation 4.

$$e_{ij} = \frac{o_{i*} \times o_{*j}}{o_{**}} \tag{3}$$

$$G^2 = 2\sum_{i,j} o_{ij} \times \log\frac{o_{ij}}{e_{ij}} \tag{4}$$

Relevance feedback

The second method for identifying key terms is based on relevance feedback in Information Retrieval,[38] which has previously been used to generate labeled training data for WSD.[39 11] Given a set of documents, D, we assume that some are labeled with a MeSH code, D_+, while the remainder, D_- are not. The number of times a term, t, occurs in a document, d in D, is represented as count (t; d) and the number of documents containing t in D as df(t). A score indicating the association between the term t and D_+ can then be computed using equation 5. In this equation, idf (t) is the inverse document frequency (IDF) of the term t and is computed using equation 6, where df(t) is the number of documents in D that contain t. IDF is a commonly used measure in Information Retrieval which provides information about the number of documents in which a term appears and assigns high values to terms that appear infrequently. When relevance feedback is used in Information Retrieval, the idf term in equation 5 is not normally included. However, when it is being used to find terms that are indicative of a domain, it is helpful to include it, since not doing so leads to infrequent terms being scored highly.[30]

$$\text{score}(t, D_+) = \text{idf}(t) \times \left(\frac{\sum_{\forall d \in D_+} \text{count}(t,d)}{|D_+|} - \frac{\sum_{\forall d \in D_-} \text{count}(t,d)}{|D_-|} \right) \tag{5}$$

$$\text{idf}(t) = \log\frac{|D|}{\text{df}(t)} \tag{6}$$

Table 1 Contingency table showing distribution of terms in documents

	Medical Subject Heading code		
	+	−	Totals
Term			
+	o_{++}	o_{+-}	o_{+*}
−	o_{-+}	o_{--}	o_{-*}
Totals	o_{*+}	o_{*-}	o_{**}

Identifying key terms for NLM-WSD

A set of key terms were generated for the MeSH codes associated with the abstracts in the NLM-WSD corpus. For each of these MeSH codes, 100 abstracts were downloaded from Medline using Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). The downloaded abstracts were then run through MetaMap to identify the candidates for each term. (MetaMap also identifies compound terms.) The processed corpus is then analyzed using the lexical statistics, and the top 10 key terms extracted for each MeSH code. For example, the key terms for the MeSH code Cultured Cells include 'cells,' 'inhibitors,' 'virus,' 'carcinoma cell,' and 'human cells.'

The key terms are used as context for each abstract that is annotated with that MeSH term. Abstracts in the NLM-WSD are typically labeled with several MeSH codes, and the context is created by taking the combination of all keys terms for the MeSH codes that apply to an abstract.

MeSH codes are not evenly distributed in Medline abstracts. On average, MeSH codes apply to 6.1 abstracts in the NLM-WSD corpus, but the most common code (Human) is associated with 2624 (76.6% of the total). This is taken into account using the IDF of each MeSH code. This is a different application of the IDF measure from that used to compute the importance of terms when computing the domain context. IDF of MeSH codes is computed using equation 7, where m is a MeSH code, M the corpus of abstracts downloaded from Medline, and df(m) the number of abstracts in M that have the MeSH code m associated with them.

$$\text{idf}(m) = \log\frac{|M|}{\text{df}(m)} \tag{7}$$

When IDF weighting is applied, the key terms for each MeSH code are weighted by the IDF score for that code, thereby reducing the importance of very common codes, such as Human, Male, or Female.

Example key terms and contexts

Table 2 shows some of the key terms that are identified by the relevance feedback method for two MeSH terms that tend to be associated with different meanings of the term 'culture.' The MeSH term Cells, Cultured is often associated with abstracts in which 'culture' is used to mean 'laboratory procedure,' while occurrences in abstracts labeled with Socieconomic Factors 'culture' almost always means 'anthropological culture.' It can be seen that there are clear differences between the terms that are identified for each MeSH code and those that would intuitively be expected to be associated with the different meanings of 'culture.'

Table 2 Example key terms identified by relevance feedback approach for Medical Subject Heading codes associated with different meanings of 'culture'

Cells, cultured	Socio-economic factors
Inhibitors	Health
Cell	Education
Virus	Income
Inhibition	Social
Assay	Countries
Inhibited	Economic
Cytotoxicity	Care
Staining	Need
Virions	Services
Epithelial	Children

Table 3 shows example contexts for a sentence from the NLM-WSD corpus which contains the word 'culture' to mean 'laboratory procedure.' The row labeled 'Local context' shows the context created from the terms found around the ambiguous word. All context terms are assigned the default weight of 1 (indicated by #1). The next row, 'Key terms,' shows context created from the MeSH term Cells, Cultured which is associated with the abstract in which this sentence appears. The next row, 'Key terms (IDF),' shows the same context with IDF weighting. In this case, all context terms are weighted 1.36, the IDF score for the MeSH term Cells, Cultured. When these are combined with the local context (bottom row), context terms are assigned different weights.

RESULTS AND DISCUSSION

Results are shown in table 4. (Performance figures for local context reported here are slightly higher than those previously reported.[20] The difference was caused by the use of a newer version (0.1.6) of the PPR software, which fixed several minor bugs.) Performance is measured as the percentage of instances correctly disambiguated. The column 'Local context' shows performance when the context around the ambiguous word is used, and corresponds to the use of PPR over the UMLS Metathesarus without any additional information about the domain of the documents. Results are also reported when each of the contexts created using the domain is used, both alone and in combination with the local context. G^2 indicates that the context is generated using the log-likelihood score and RF that the relevance feedback approach was used. Both methods are applied with and without IDF scores of the MeSH terms being used to weight the context (indicated by 'IDF' in table 4 when it is used). Results obtained using the domain model are compared with the local context and statistical significance computed using bootstrap resampling with 95% confidence.[40]

Table 3 Samples of contexts generated for the sentence 'The main goal of the present study was to determine whether or not oligodendrocytes in culture constitutively express the different βAPP isoforms' (simplified for brevity).

Local context	goal#1 present#1 study#1 oligodendrocytes#1 culture#1 different#1 isoforms#1
Key terms	inhibitors#1 cell#1 virus#1 inhibition#1 assay#1
Key terms (IDF)	inhibitors#1.36 cell#1.36 virus#1.36 inhibition#1.36 assay#1.36
Local context and Key terms (IDF)	goal#1 present#1 study#1 oligodendrocytes#1 culture#1 different#1 isoforms#1 inhibitors#1.36 cell#1.36 virus#1.36 inhibition#1.36 assay#1.36

IDF, inverse document frequency.

The results for local context and domain models are comparable. However, when the local and domain contexts are combined, performance is significantly better than when only local context is used. This indicates that terms from the domain contain useful information for WSD that is different from the local context. Note also that when the domains model is used alone, disambiguation is performed at the MeSH code level: words in abstracts labeled with the same MeSH codes are tagged with the same sense. However, when adding local contexts, the system is able to discriminate among contexts, thus providing a more fine-grained disambiguation. Performance consistently improves when IDF weighting is used. Improvement is observed regardless of which method is used to generate the domain context and whether the domain context is used alone or in combination with local context. This improvement shows that applying the IDF weighting is able to accommodate the skewed assignment of MeSH codes to abstracts. The relevance feedback method for generating domain context produces higher results than the log likelihood, although the difference is not significant.

Note that our results compare favorably to all knowledge-based systems reported in Jimeno-Yepes and Aronson,[11] which reports a best result of 68.36. In order to compare the results with McInnes,[16] the second row of table 4 reports our results for the same subset of 13 terms. The relevance feedback method with IDF weighting significantly also outperforms the local context for these terms and is over 12 points higher than the performance reported by McInnes[16] (48.1).

Results are also reported for each term in the NLM-WSD data set. The column labeled 'count' shows the number of instances of each term that were used for the experiments. There is a wide variation in performance over individual terms. Disambiguation for some terms (eg, fat, pressure, secretion, surgery, and transient) is very high with near-perfect disambiguation. However, for other terms (such as fit and inhibition), performance is very poor. The use of domain information leads to a large improvement in performance for many terms, and in general, the improvement is observed regardless of which domain model is used. For example, results for the term man increase from 45.7 to between 76.9 and 87.0 depending on the domain model. Other terms for which the domain models lead to substantial increases in performance include cold, extraction, nutrition, reduction, and sensitivity.

Although the overall performance improves when the domain model is used, there are some terms for which the results get worse. The drop in performance for the term condition is particularly striking. This term has two possible meanings in the NLM-WSD corpus: 'a state of being, such as state of health' and 'psychological conditioning.' The first meaning applies to 90 of the 92 instances of condition in the NLM-WSD corpus and is quite general, which leads to the low performance of the approach using local context alone. However, some abstracts in which this meaning appears also contain MeSH terms that lead to the second sense being preferred through connections in the graph created from the UMLS—for example, Anxiety Disorder, Behavior, and Depressive Disorder.

Our approach relies on converting the UMLS Metathesaurus into a graph and computing the contexts associated with each-MeSH code. Although these tasks are time-consuming, they are typically performed off line. When applying our system to text, it has to be preprocessed with MetaMap and then run through PPR. Overall, the system is able to disambiguate 37 instances per minute on a PC with 2 QuadCore Xeon3 160 Mhz processors and 32 GB of RAM.[20] The runtime overhead of augmenting local context with domain information is negligible.

Table 4 Word Sense Disambiguation results using local and domain context

			Domain context alone				Domain context and local context			
	Count	Local context	G^2	G^2 (IDF)	RF	RF (IDF)	G^2	G^2 (IDF)	RF	RF (IDF)
All		70.4	70.0	70.8	70.6	71.5	72.8	73.5*	73.5*	**73.7***
McInnes subset		54.5	57.5	57.9	57.7	58.6	58.9	59.2	59.1	**60.2**
Adjustment	93	33.3	32.3	34.4	35.5	33.3	34.4	35.5	38.7	37.6
Blood pressure	100	46.0	51.0	52.0	52.0	53.0	52.0	52.0	53.0	54.0
Cold	95	30.5	60.0	64.2	63.2	64.2	66.3	67.4	68.4	68.4
Condition	92	41.3	6.5	15.2	8.7	5.4	13.0	20.7	13.0	9.8
Culture	100	80.0	87.0	91.0	83.0	86.0	88.0	92.0	85.0	86.0
Degree	65	92.3	95.4	93.8	96.9	96.9	95.4	95.4	95.4	95.4
Depression	85	88.2	100.0	98.8	98.8	98.8	98.8	97.6	98.8	97.6
Determination	79	94.9	73.1	87.2	87.2	84.6	79.7	91.1	87.3	83.5
Discharge	75	81.3	82.7	81.3	84.0	82.7	85.3	84.0	89.3	84.0
Energy	100	95.0	86.9	86.9	93.9	92.9	87.0	87.0	94.0	93.0
Evaluation	100	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Extraction	87	28.7	32.2	39.1	33.3	40.2	35.6	41.4	35.6	42.5
Failure	29	93.1	86.2	82.8	65.5	79.3	86.2	86.2	75.9	82.8
Fat	73	95.9	97.3	97.3	97.3	97.3	97.3	97.3	97.3	97.3
Fit	18	11.1	5.6	5.6	0.0	0.0	11.1	11.1	5.6	5.6
Fluid	100	90.0	92.9	92.9	93.9	93.9	90.0	90.0	91.0	92.0
Frequency	94	98.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ganglion	100	73.0	69.0	72.0	71.0	73.0	80.0	79.0	80.0	81.0
Glucose	100	90.0	92.9	92.9	91.9	91.9	92.0	94.0	92.0	92.0
Growth	100	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0
Immunosuppression	100	62.0	73.0	73.0	74.0	74.0	73.0	74.0	74.0	74.0
Implantation	98	87.8	70.4	83.7	74.5	88.8	76.5	90.8	84.7	93.9
Inhibition	99	3.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0
Japanese	79	81.0	82.3	78.5	84.8	82.3	84.8	79.7	86.1	82.3
Lead	29	93.1	20.7	20.7	93.1	93.1	93.1	93.1	93.1	93.1
Man	92	45.7	81.3	85.7	76.9	82.4	84.8	87.0	81.5	83.7
Mole	84	57.1	56.6	53.0	62.7	57.8	69.0	65.5	72.6	70.2
Mosaic	97	71.1	59.8	56.7	59.8	58.8	67.0	67.0	70.1	71.1
Nutrition	89	29.2	49.4	53.9	46.1	52.8	47.2	50.6	44.9	49.4
Pathology	99	33.3	16.7	17.7	16.7	17.7	20.2	22.2	22.2	22.2
Pressure	96	97.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Radiation	98	52.0	43.9	43.9	43.9	43.9	43.9	43.9	42.9	43.9
Reduction	11	45.5	72.7	72.7	72.7	72.7	72.7	72.7	72.7	63.6
Repair	68	76.5	80.9	80.9	79.4	82.4	82.4	82.4	79.4	82.4
Resistance	3	66.7	100.0	100.0	100.0	100.0	66.7	66.7	66.7	66.7
Scale	65	81.5	82.8	81.2	82.8	82.8	73.8	73.8	72.3	75.4
Secretion	100	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0
Sensitivity	51	33.3	62.7	62.7	62.7	62.7	64.7	62.7	64.7	64.7
Sex	100	87.0	85.0	82.0	85.0	83.0	86.0	84.0	86.0	85.0
Single	100	94.0	87.0	86.0	79.0	85.0	91.0	89.0	90.0	90.0
Strains	93	94.6	91.4	86.0	95.7	90.3	96.8	95.7	96.8	95.7
Support	10	90.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0
Surgery	100	97.0	98.0	98.0	98.0	98.0	98.0	98.0	98.0	98.0
Transient	100	99.0	92.9	97.0	88.9	96.0	98.0	99.0	98.0	99.0
Transport	94	95.7	98.9	98.9	98.9	98.9	97.9	97.9	97.9	97.9
Ultrasound	100	83.0	84.0	84.0	82.0	82.0	84.0	84.0	84.0	82.0
Variation	100	90.0	83.0	67.0	73.0	70.0	88.0	81.0	88.0	83.0
Weight	53	60.4	56.6	60.4	60.4	60.4	60.4	56.6	64.2	64.2
White	90	60.0	58.9	62.2	58.9	63.3	71.1	71.1	71.1	73.3

*Statistical significance with respect to the local context baseline, computed using bootstrap resampling.[40]
Terms used in McInnes subset[16] are shown in italics.
IDF, inverse document frequency.

CONCLUSIONS

This paper demonstrates that including information about the domain in which ambiguous words appear significantly improves the performance of a knowledge-based WSD algorithm for medical documents,[20] and over other knowledge-based systems.[11 16] Domain information has already been shown to provide useful information for WSD in general,[21–23] and the results reported here show that it also improves WSD performance for documents that share related topics, such as medical texts.

The WSD system described here uses a knowledge-based approach. It has the advantage of not requiring labeled training data (unlike several other studies[7–9 25–28]) and being able to distinguish between UMLS concepts with the same Semantic Type (unlike Humphrey *et al*[13]). The system was evaluated on

the set of terms in the NLM-WSD corpus but could disambiguate all words in a document that are ambiguous in the UMLS Metathesaurus. A novel method for representing domain information was also introduced in which the domain is represented as a set of key terms. These are used as context for the WSD algorithm, either alone or in combination with local context. The WSD algorithm used is a 'bag of words' model, which does not make use of information about the order in which terms occur in the context of ambiguous words, and can make use of the key terms extracted for each domain in a straightforward way. Key terms are identified by applying lexical statistics to a corpus in which documents are labeled with domain code. Two lexical statistics were explored, and it was found that one based on relevance feedback provided the best performance. The frequency of domain labels was also found to be important, and the IDF statistic was used to weight key terms and to reduce the importance of those which occur frequently.

The approach for identifying key terms described in the paper assumes that a corpus with domain labels is available. The MeSH codes in Medline provide suitable domain labels that have been assigned by human annotators and are therefore likely to be accurate. Alternative methods could be used to generate domain labels if manually annotated labels were not available. For example, labels could be assigned automatically using text categorization,[41 42] or examples of documents on a particular topic can be gathered automatically.[43] We plan to explore these alternative methods for generating domain labels in future work. In addition, we also plan to explore performance on other genres of medical documents, such as clinical texts.[44]

Funding MS is grateful for support from the Engineering and Physical Sciences Research Council (EP/D069548/1). EA and AS are grateful for support from the Ministry of Science (KNOW2—TIN2009-14715-C04-01): (1) Ministerio de Educacion y Ciencia; (2) Engineering and Physical Sciences Research Council.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Shatkay H.** Hairpins in book stacks: information retrieval from biomedical text. *Brief Bioinform* 2005;**6**:222—38.
2. **Krallinger M,** Valencia A. Text mining and information retrieval services for molecular biology. *Genome Biol* 2005;**6**:224.
3. **Westbrook JI,** Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? *J Am Med Inform Assoc* 2005;**12**:315—21.
4. **Lau AY,** Coiera EW. Impact of web searching and social feedback on consumer decision making: a prospective online experiment. *J Med Internet Res* 2008;**10**:e2.
5. **Ide N,** Véronis J. Introduction to the special issue on word sense disambiguation: the state of the art. *Comput Ling* 1998;**24**:1—40.
6. **Navigli R.** Word sense disambiguation: a survey. *ACM Comput Surv* 2009;**41**:1—69.
7. **Joshi M,** Pedersen T, Maclin R. A comparative study of support vector machines applied to the word sense disambiguation problem for the medical domain. *Proceedings of the Second Indian Conference on Artificial Intelligence (IICAI-05)*. Pune, India, 2005:3449—68.
8. **Savova GK,** Coden AR, Sominsky IL, *et al*. Word sense disambiguation across two domains: biomedical literature and clinical notes. *J Biomed Inform* 2008;**41**:1088—100.
9. **Liu H,** Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc* 2004;**11**:320—31.
10. **Artstein R,** Poesio M. Inter-coder agreement for computational linguistics. *Comput Ling* 2008;**34**:555—96.
11. **Jimeno-Yepes AJ,** Aronson AR. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics* 2010;**11**:569.
12. **Stevenson M,** Guo Y. Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus. *J Biomed Inform* 2010;**43**:762—73.
13. **Humphrey SM,** Rogers WJ, Kilicoglu H, *et al*. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. *J Am Soc Inf Sci Technol* 2006;**57**:96—113.
14. **Humphrey SM.** Automatic indexing of documents from journal descriptors: a preliminary investigation. *J Am Soc Inf Sci* 1999;**50**:661—74.
15. **Nelson S,** Powell T, Humphreys B. The Unified Medical Language System (UMLS) project. In: Kent A, Hall CM, eds. *Encyclopedia of Library and Information Science*. New York: Marcel Dekker, Inc, 2002.
16. **McInnes B.** An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and medline. *Proceedings of the ACL-08: HLT Student Research Workshop*. Columbus, Ohio, 2008:49—54.
17. **Lesk M.** Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of ACM SIGDOC Conference*. Toronto, Canada, 1986:24—6.
18. **Navigli R,** Lapata M. Graph connectivity measures for unsupervised word sense disambiguation. *Proceedings of IJCAI*. Hyderabad, India, 2007:1683—8.
19. **Agirre E,** Soroa A. Personalizing PageRank for word sense disambiguation. *Proceedings of EACL-09*. Athens, Greece, 2009.
20. **Agirre E,** Soroa A, Stevenson M. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics* 2010;**26**:2889—96.
21. **Koeling R,** McCarthy D, Carroll J. Domain-specific sense distributions and predominant sense acquisition. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. HLT/EMNLP*. Ann Arbor, MI, 2005:419—26.
22. **Agirre E,** de Lacalle OL, Soroa A. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009:1501—6.
23. **Khapra M,** Kulkarni A, Sohoney S, *et al*. All words domain adapted WSD: finding a middle ground between supervision and unsupervision. *Proceedings of ACL 2010*. Uppsala, Sweden, 2010:1532—41. http://www.aclweb.org/anthology/P10-1155.
24. **Yu H,** Kim W, Hatzivassiloglou V, *et al*. A large scale, corpus-based approach for automatically disambiguating biomedical abbreviations. *ACM Trans Inform Syst* 2006;**24**:380—404.
25. **Xu H,** Fan JW, Hripcsak G, *et al*. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics* 2007;**23**:1015—22.
26. **Stevenson M,** Guo Y, Gaizauskas R, *et al*. Disambiguation of biomedical text using a variety of knowledge sources. *BMC Bioinformatics* 2008;**9**(Suppl 11):S7.
27. **Stevenson M,** Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. *J Biomed Inform* 2010;**46**:972—81.
28. **Jimeno Yepes A,** Aronson A. Self-training and co-training in biomedical word sense disambiguation. *Proceedings of BioNLP 2011 Workshop*. Portland, OR: ACL, 2011:182—3.
29. **Brin S,** Page L. The anatomy of a large-scale hypertextual Web search engine. *Comput Netw ISDN Syst* 1998;**30**:107—17.
30. **Haveliwala TH.** Topic-sensitive PageRank. *WWW'02: Proceedings of the 11th International Conference on World Wide Web*. New York: ACM, 2002:517—26.
31. **Reddy S,** Inumella A, McCarthy D, *et al*. Domain specific word sense disambiguation. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: ACL, 2010.
32. **Aronson AR,** Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229—36.
33. **Weeber M,** Mork J, Aronson A. Developing a test collection for biomedical word sense disambiguation. *Proceedings of AMIA Symposium*. Washington, DC: AMIA, 2001:746—50.
34. **Dunning T.** Accurate methods for computing the statistics of surprise and coincidence. *Comput Ling* 1993;**19**:61—74.
35. **Pedersen T.** A decision tree of bigrams is an accurate predictor of word sense. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*. Pittsburgh, PA: ACL, 2001:79—86.
36. **Korkontzelos I,** Manandhar S. Detecting compositionality in multi-word expressions. *Proceedings of the ACLIJCNLP 2009 Conference Short Papers*. Suntec, Singapore, 2009:65—8. http://www.aclweb.org/anthology/P/P09/P09-2017.
37. **Rayson P,** Garside R. Comparing corpora using frequency profiling. *The Workshop on Comparing Corpora*. Hong Kong, China, 2000:1—6. http://www.aclweb.org/anthology/W00-0901.
38. **Rocchio J.** Relevance feedback in information retrieval. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice Hall Inc, 1971.
39. **Stevenson M,** Guo Y, Gaizauskas R. Acquiring sense tagged examples using relevance feedback. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*. Manchester, UK: COLING, 2008.
40. **Noreen E.** *Computer-Intensive Methods for Testing Hypotheses*. New York: John Wiley & Sons, 1989.
41. **Manning H,** Schütze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.
42. **Sebastiani F.** Machine learning in automated text categorization. *ACM Comput Surv* 2002;**34**:1—47.
43. **Kilgarriff A,** Reddy S, Pomikalek J, *et al*. A corpus factory for many languages. *LREC Workshop on Web Services and Processing Pipelines*. Valetta, Malta: ELRA, 2010.
44. **Roberts A,** Gaizauskas R, Hepple M, *et al*. Semantic annotation of clinical text: the CLEF corpus. *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*. Marrakech: ELRA, 2008:19—26.