Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods

Rachel Chasin,[1] Anna Rumshisky,[2] Ozlem Uzuner,[3] Peter Szolovits[1]

[1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Department of Computer Science, University of Massachusetts, Lowell, Massachusetts, USA
[3]Department of Information Studies, University at Albany, SUNY, Albany, New York, USA

Correspondence to
Professor Anna Rumshisky, Department of Computer Science, University of Massachusetts Lowell, 198 Riverside St, Olsen Hall 215, Lowell, MA 01854, USA; arum@cs.uml.edu

Received 27 June 2013
Revised 18 December 2013
Accepted 23 December 2013

ABSTRACT

Objective To evaluate state-of-the-art unsupervised methods on the word sense disambiguation (WSD) task in the clinical domain. In particular, to compare graph-based approaches relying on a clinical knowledge base with bottom-up topic-modeling-based approaches. We investigate several enhancements to the topic-modeling techniques that use domain-specific knowledge sources.

Materials and methods The graph-based methods use variations of PageRank and distance-based similarity metrics, operating over the Unified Medical Language System (UMLS). Topic-modeling methods use unlabeled data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database to derive models for each ambiguous word. We investigate the impact of using different linguistic features for topic models, including UMLS-based and syntactic features. We use a sense-tagged clinical dataset from the Mayo Clinic for evaluation.

Results The topic-modeling methods achieve 66.9% accuracy on a subset of the Mayo Clinic's data, while the graph-based methods only reach the 40–50% range, with a most-frequent-sense baseline of 56.5%. Features derived from the UMLS semantic type and concept hierarchies do not produce a gain over bag-of-words features in the topic models, but identifying phrases from UMLS and using syntax does help.

Discussion Although topic models outperform graph-based methods, semantic features derived from the UMLS prove too noisy to improve performance beyond bag-of-words.

Conclusions Topic modeling for WSD provides superior results in the clinical domain; however, integration of knowledge remains to be effectively exploited.

INTRODUCTION

The past decade has seen a surge of interest in data mining and information extraction over clinical text such as admission notes, nurses' notes, and discharge summaries. Despite the pervasive presence of domain-specific lexical ambiguity in clinical text, which significantly impedes such efforts, there has been a lack of a unified concerted effort to address this problem. In this paper, we address this methodological gap by conducting an evaluation of some of the most promising word sense disambiguation/induction (WSD/WSI) methods that have been developed for the general-domain English text as well as for the domain of biomedical literature.

While there is a large body of research on WSD in general English and some recent efforts in the biomedical literature domain, lexical ambiguity in clinical text has received much less attention. While some of the principles and methods may translate well between different domains, such success is not guaranteed with techniques that use domain-specific knowledge resources or text-processing tools that need to be trained on domain-specific text. Clinical prose is remarkably different from both general-domain English and biomedical text. In the clinical domain, institution-specific templates, abbreviation conventions, and non-standard sentence structure and phrasing create additional difficulties for information extraction and reasoning over text.[1] The following examples from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database, which contains patient records from intensive care units at the Beth Israel Deaconess Medical Center,[2] illustrate some of the typical problems, including lexical ambiguity, in particular:

'PAIN: MEDICATED WITH TOTAL 4 MG IV MS, C/O INCISIONAL PAIN, SOME RELIEF WITH IV MS'

'slides down freq. in bed or chair. ms- confused, cooperative. poor short term memory'.

In the first example, the acronym 'ms' corresponds to 'morphine sulfate', while in the second example, 'ms' should be interpreted as 'mental state'.

Any task that uses machine learning methods to extract information from clinical text, such as automatic cohort selection and compilation of disease presentations, stands to benefit directly from a more accurate representation of relevant text. The latter includes both the disambiguation of lexical ambiguities and the related task of abbreviation/acronym expansion. The largest barrier to accurate WSD is the cost of annotating data used by supervised learning methods. Clinical text annotation must be performed by medical experts, and creating an annotated dataset for every ambiguous word is prohibitively expensive. Many efforts in WSD therefore focus on unsupervised or semi-supervised methods, requiring few or no annotated data, while also attempting to leverage expert-curated knowledge bases in order to incorporate human expertise.

One disadvantage of unsupervised word sense induction, in which sense clusters are induced purely from unlabeled data, is that these clusters of examples are not mapped to an existing inventory of senses. Whether or not this mapping is necessary depends on the practical applications; for further implications of this, see the discussion section below. In this paper, we compare the graph-based

To cite: Chasin R, Rumshisky A, Uzuner O, *et al*. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2013-002133

WSD methods that use the Unified Medical Language System (UMLS) with the bottom-up induction techniques that adopt a Bayesian topic-modeling approach to the problem of lexical ambiguity. We show that the latter provide superior results even when only simple bag-of-words (BOW) features are used. We further investigate the impact of incorporating syntactic and knowledge-based features into the topic model. We evaluate the performance of these approaches on ambiguous clinical terms from Mayo Clinical Corpus (MCC), a sense-tagged dataset from the Mayo Clinic.[3]

The rest of this paper is organized as follows. In the next section, we give an overview of related work. We then describe the data used in experiments, followed by a presentation of the methods used in both graph-based and Bayesian topic-modeling experiments. We then discuss the results for each set of experiments. We conclude with some discussion of the implications and future directions for this work.

RELATED WORK

The UMLS,[4] which is the dominant knowledge source in the biomedical domain, assigns a unique identifier (CUI) to each medical concept. UMLS maps strings to their possible meanings (CUIs) and connects CUIs to each other with relations such as 'broader than' and 'narrower than'. It also assigns each CUI to a 'semantic type', a broad category such as 'Finding' or 'Disease or Syndrome'. This information is largely sourced from other medical vocabularies.

One widely used application that processes clinical text is MetaMap,[5] which includes an optional WSD step[6] that disambiguates mainly at the semantic-type level using statistical associations between words and 'Journal Descriptors'.[7]

The UMLS has been widely used for WSD in the biomedical domain.[8–10] When a knowledge base or an ontology is used for WSD, in both the general domain[11] and the biomedical domain,[8 10 12] it has been treated as a graph whose nodes are concepts (CUIs in UMLS) and whose edges are relations between them. Graph-based methods that derive relative ranking of UMLS nodes corresponding to senses have been found to outperform other approaches.[10] Agirre *et al*[8] have run a variant of PageRank[13] over this graph to distribute weight over CUIs and pick the target's CUI with the most weight. Other work[9] restricts UMLS to a tree and uses tree similarity measures to assign scores to CUIs of the target based on CUIs of context words. All approaches that use the graph-like properties of UMLS are susceptible to shortcomings in UMLS's structure, and tend to improperly favor senses that are more connected and thus more easily reachable.

Clustering has also been applied to WSD in the general domain and beyond.[14] One of the challenges in this is to determine the number of clusters to create—that is, the stopping condition for the clustering. Savova *et al*[15] investigated this on biomedical text. One of the recent evaluations of the state of the art in word sense induction in the general domain was conducted at SemEval-2010,[16] where the top-performing systems achieved an accuracy of 62% using supervised evaluation. The participant systems focused on a variety of WSI improvements including feature-selection/dimensionality-reduction techniques,[17] experiments with bigram and co-occurrence features[17] and syntactic features,[18] and increased scalability.[19] Clustering over word co-occurrence graphs[20] and second-order co-occurrence vectors[17] was used by the top-performing systems, with measures such the Gap statistic[21] used to predict the number of clusters.[17]

Supervised machine learning methods have been tested in both clinical and biomedical domains. Savova *et al*[3] experimented on the biomedical dataset from the National Library of Medicine containing Medline abstracts, as well as the Mayo WSD dataset containing clinical text, beating the most-frequent-sense (MFS) baseline. The task of abbreviation/acronym expansion has attracted some recent attention in the clinical domain.[22 23] Moon *et al*[24] conducted experiments aiming to determine a good window for BOW features, a good supervised classifier type, and the minimum number of instances needed to achieve satisfactory performance.

In the general domain, Brody and Lapata[25] have adapted a Bayesian topic-modeling method, latent Dirichlet allocation (LDA),[26] to WSI by treating each occurrence context of an ambiguous word as a document, and the derived topics as sense-selecting context patterns represented as collections of features. Yao and Van Durme[27] followed their work by applying to this task a non-parametric generative model, the hierarchical Dirichlet process (HDP).[28] The advantages of HDP over LDA lie in its ability to avoid tuning the number of clusters to create by modeling new cluster creation in addition to cluster selection as part of the algorithm. Both LDA- and HDP-based models outperform the systems that competed in the WSI task for nouns in the general domain in the 2007 SemEval competition.[29]

DATA

Mayo Clinical Corpus

MCC[3] consists of 50 ambiguous clinical term targets; 48 have 100 instances and two have 1000 instances. Each instance contains a sentence or two of context and a manually assigned CUI representing the sense of the target or 'none' if there is no such CUI.

We remove the 140 instances labeled 'none' in our experiments. The mean number of CUIs used to label each target is 4.7, with SD 2.2. The average κ value (Cohen's κ) across all targets in MCC is 0.54, representing moderate agreement.[3] We refer the reader to the original paper for the discussion of the agreement rate on MCC. For topic-modeling experiments, we split MCC into a mapping set (70%) and a test set (30%). We created these sets with similar proportions of senses in each, but chose instances randomly among those labeled with the same sense. The mapping set is used in evaluation and also in cross-validation to tune topic-modeling parameters and select feature types. In the graph-based experiments, we use a subset of 15 out of 50 targets, for which all of the CUIs assigned to them in the labeled data appear in Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), one of the UMLS source vocabularies. SNOMED CT provides a path up its hierarchy for each CUI, simplifying the calculations necessary for these methods.

Training data

We obtain unlabeled training data for the topic-modeling algorithms from nurses' notes and discharge summaries in the MIMIC II database. At the time of data extraction, MIMIC contained 27 059 deidentified patient records with multiple notes per record. Instances are collected for each target by comparing the targets with whitespace-delimited tokens with punctuation removed in MIMIC; if a target matches, an instance is created from the surrounding 100 tokens. Instances that overlap in content are allowed. We collect up to 50 000 instances per target, reduced if fewer instances are available. In the resulting dataset, four targets have fewer than 1000 instances (232 being the fewest), 26 targets have between 1000 and 10 000 instances, and 20 targets have over 10 000 instances. The mean number of instances is 16 026. The collection process is case-insensitive, except for the abbreviations, such as 'it', which can be a

pronoun or may have a specialized sense (eg, 'intrathecal'). For the abbreviations that have a general English sense, we only consider uppercase matches.

Data preprocessing

The generation of features to use in the topic models requires some preprocessing of the data. For each instance, we tokenize the text, find sentence boundaries, assign part-of-speech (POS) tags to the tokens, and perform dependency parsing. Dependency parsing identifies binary asymmetric dependency relations between lexical items, and is typically more robust to noise.

We use the POS tagger and dependency parser from ClearNLP,[30] which provides models trained on clinical text. We also identify the clinical phrases (CPs) present in the text. Each string of up to six tokens is normalized using the Lexical Variant Generation program (LVG), which tokenizes, stems, and alphabetizes the resulting tokens.[4] The normalized string is looked up in the English normalized string table (mrxns_eng) provided by the UMLS, and, if present, that string is considered a CP. Each token is assigned to the longest CP that it belongs to, if any. This method allows us to capture phrases such as 'l4 vertebral bodies', although it can also create false positives when stopwords such as 'is' appear in UMLS.

METHODS

Graph-based methods using the UMLS

Path-based

We perform path-based experiments using the methods of McInnes *et al*[9] in which we use UMLS as a tree whose nodes are the CUIs and whose edges are a subset of the relations (only broader/narrower and parent/child relations). A target word, t, is disambiguated to the CUI, c*, with the largest cumulative similarity to the context words according to the following general formula:

$$c^* = \underset{c \in cuis(t)}{\operatorname{argmax}} \sum_{w \in context(t)} (weight(w, t) \max_{n \in cuis(w)} similarity(c, n)).$$

The function similarity represents the similarity between nodes c and n in UMLS measured by a tree distance metric, and the function weight represents how important this context word should be in the calculation. Both of these functions may be varied, but we use a uniform weight function and the similarity function, wup,[31] which depends on the depth of each node, and their lcs (least common subsumer), the deepest node that is an ancestor of both.

$$wup(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

PageRank-based

We use the methods of Agirre *et al*,[8] which involve variants on the PageRank algorithm.[13] In these methods, PageRank is run on a graph whose nodes are CUIs in UMLS chosen on the basis of the target's context and whose edges are relations present between them. After PageRank is run, the target is disambiguated to the most 'popular' CUI—that is, the one with the most weight. Intuitively, each sense of each context word carries weight that can be spread among different nodes in the UMLS graph. The senses of the target word that are similar to the senses of the context word (ie, closer to it in the graph) should receive more of this weight.

The UMLS graph can be altered on the basis of the context around the target word in the following ways: (1) create a subgraph based on the context and run traditional PageRank (SPR); (2) use the whole graph but run PageRank with a non-uniform ('personalized') initial weight vector (PPR). In SPR, the subgraph is created by identifying the nodes associated with all context CPs, finding the shortest paths in UMLS between each pair of these nodes, and including exactly the nodes on those paths. In our experiments, we approximate the shortest path between two nodes as the concatenation of their paths to their least common subsumer, treating UMLS as a tree. This would be the shortest path in an actual tree, and we use this approximation to keep our experiments feasible. In PPR, all of UMLS is used as the graph, but the initial vector is weighted so that only nodes associated with context CPs have non-zero weights. Although traditional PageRank's initial vector is uniformly weighted over the graph, we also experiment with distance-based weighting, where each node's weight is inversely proportional to its distance from the target in tokens. We use 40 iterations of PageRank for SPR, and 20 for PPR. Our results on clinical text suggest that that PPR is more effective than SPR, supporting the findings of Agirre *et al*[8] in the biomedical domain.

Bayesian topic-modeling-based methods

Models

LDA is a Bayesian topic-modeling technique which is more formally defined as follows. Consider M instances of a target word that has K 'topics'—that is, sense-selecting context patterns. Let the context of instance j be described by some set of Nj features from a vocabulary of size V. LDA assumes that there are M probability distributions $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots \theta_{jK})$, where θ_{jk} = the probability of generating topic k for instance j, and K probability distributions $\phi_k = (\phi_{k1}, \phi_{k2}, \ldots \phi_{kV})$, where ϕ_{kf} = the probability of generating feature f from topic k. This makes the probability of generating the corpus where the features for instance j are $f_{j1}, f_{j2}, \ldots, f_{jN_j}$

$$P(corpus) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \sum_{k=1}^{K} \theta_{jk} \phi_{kf_{ji}}$$

Figure 1 shows the plate representation of LDA (compare with Blei *et al*[26]).

The goal of LDA for WSI is to obtain the distribution $\boldsymbol{\theta}_{j*}$ for an instance j* of interest, as this corresponds to the probability of being the correct sense for the target word in this context.

The corpus generation process for HDP is similar to that of LDA, but obtains the document-specific sense distribution via a Dirichlet process whose base distribution is determined via another Dirichlet process, allowing for an unfixed number of topics because the draws from the resulting topic distribution are not limited to a preset range. The concentration parameters of both Dirichlet processes are determined via hyperparameters.

We train a separate model with its own set of topics for each target. In order to reduce the effects of randomization during

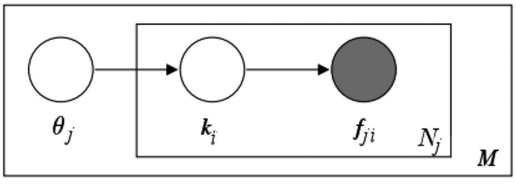

Figure 1 Graphical model representation of latent Dirichlet allocation. The outer plate represents each of the M documents. The inner plate represents each of the N_j features in the jth document. Each feature is assigned a topic from a document-specific distribution (θ_j).

Gibbs sampling in the LDA implementation we run, we average results over five models trained for each target using the same parameters and data. The way we average results is described below in the section on evaluation metrics. While randomization is present in the inference algorithms, we perform only one inference run per model.

Our process for training an LDA model uses the software, GibbsLDA++,[32] which uses Gibbs sampling to assign topics to each feature in each instance. We use the hyperparameters used by Brody and Lapata,[25] who tuned the value for α and used a known value for β from previous LDA work: $\boldsymbol{\alpha}=0.02$, $\boldsymbol{\beta}=0.1$. The HDP training and inference procedures are similar to LDA, but using Gibbs sampling on topic and table assignment in a Chinese restaurant process. We use Wang's program for HDP[33] using Yao and Van Durme's[27] hyperparameters $\mathbf{H}=0.1$, $\boldsymbol{\alpha}_0 \sim$ Gamma(0.1, 0.028), $\boldsymbol{\gamma} \sim$ Gamma(1, 0.1).

Feature types

In addition to the basic BOW features, we generate additional feature classes using dependency parsing and UMLS. One class is the similar 'bag-of-CPs' features—an unordered set of the closest CPs as identified during preprocessing, represented by their LVG normalizations. We also experiment with features based on ontological information about CPs syntactically connected to the target. We use CPs reachable from the target in three or fewer syntactic dependencies ('hops'). This number is selected by manual examination of important syntactic relations. When more than one hop is involved, dependency information for all hops is included, and the order of relations is preserved. Purely syntax-based features are created by prepending the stemmed token from the dependency to syntactic information.

Two types of UMLS-based features are also generated for each of the relevant CPs' possible CUIs: ancestor features and semantic-type features. We define the 'Kth ancestor' of a CUI c as all CUIs that have 'parent' (PAR) or 'broader than' (RB) relations to any of the '(K−1)th ancestors' of c, and we define the '0th ancestor' as c itself. Unlike a true tree, where each node has exactly one parent, UMLS CUIs often have many parents. Because of this high fan out, we only generate the 0th ancestor through 2nd ancestors. A feature is produced from each ancestor by prepending the ancestor's CUI to the syntactic information.

UMLS's semantic types have IDs (TUIs) and are arranged in a hierarchy, in this case a true tree, so each type has only one parent. We distinguish two feature classes generated from CUIs of context CPs: one includes the CUI's semantic type; the other also includes additional features for all of the type's ancestors. Each single feature corresponds to a TUI associated with a particular syntactic dependency found by the parser.

Figure 2 illustrates some of the feature types generated for the following instance of the target 'compression':

'4). Unchanged appearance of sclerotic metastases involving the L3 and L4 vertebral bodies, with L4 **compression** deformity. Subsequently, an MRI was performed and showed'

Evaluation metrics

Following the established practice in SemEval competitions and subsequent work,[16 25 27 29] we conduct supervised evaluation. A small amount of labeled data, the mapping set, is used to map the induced topics (corresponding to sense-selecting patterns) to real-world senses. The mapping produced is probabilistic; for topics $1,\ldots,K$ and senses $1,\ldots,S$, we compute the KS values $P(s \mid k)$ = count(instances predicted k, labeled s)/count (instances predicted k). Then, given $\boldsymbol{\theta}_{j*}$, we can make a prediction for instance j* that is better than just the most likely sense for its most likely topic. Instead, we compute $\underset{1\le s\le S}{\operatorname{argmax}} \sum_{k=1}^{K} \theta_{j*k} P(s)$, the sense with the highest probability of being correct for this instance, given the topic probabilities and the KS mapping probabilities.

The supervised evaluation measures traditionally reported in natural language processing tasks include precision, recall, F-measure, and accuracy. For WSD, these measures are defined in terms of the number of correct predictions (C), the number of total predictions (P), and the total number of instances (T). Precision is the percentage of instances with predictions that are correct —that is, C/P. Recall is the percentage of all instances that are correct —that is, C/T. F-measure is the harmonic mean of precision and recall. Accuracy is the percentage of correct instances —that is, the same as recall, C/T. Since our WSI system assigns a sense to every instance and no instances are left uncategorized, P=T, so precision, recall, F-measure, and accuracy are all equivalent, and we will report accuracy throughout this paper.

In cross-validation, the reported accuracy for a given configuration is averaged over five trained models. On the test data, we report the accuracy obtained by assigning to each instance the sense that the majority of the models predicted for it. The results we report use averages taken over all targets, since results for individual targets may vary.

RESULTS AND DISCUSSION

Graph-based results

Table 1 shows our best graph-based results on the SNOMED CT subset of MCC. We report the micro-average (in this case, the same as the macro-average) over all the targets' accuracies, as well as the micro-average 95% CI. The accuracies are 42.5% for the path-based and 48.9% for the PageRank-based method. As the table shows, these graph-based methods fail to reach the MFS baseline performance (56.5%). For comparison, we also show a result on this subset obtained by the topic-modeling approach using the best number of topics for this subset and a similar context window (66.9%).

Figure 2 Examples of feature classes generated for an instance of the target 'compression'.

Feature class	Example feature from instance
Bag-of-words (stemmed non-stopwords)	vertebr
Bag-of-CPs (6 closest CPs)	body l4 vertebral
Token-populated syntax	deformity_adjectival-modifier
Semantic-type-populated syntax	acquired_abnormality_T020_adjectival-modifier
UMLS-ancestor-populated syntax	Acquired_deformity_C0221430_adjectival-modifier

Table 1 Accuracies on the SNOMED subset of MCC for graph-based and topic-modeling experiments

Configuration	Accuracy (%)	95% CI
MFS	56.5	54 to 59
PPR, 20 closest CPs, all relations, initial vector weighted by inverse distance	48.9	46 to 51
SPR, 20 closest CPs, relations from path-to-hierarchy-root, initial vector weighted by inverse distance	43.5	41 to 46
Path, CPs in 70-word window, similarity measure wup, uniform CP weighting	42.5	40 to 45
LDA, 20 closest words, 5 topics	66.9	64 to 70

CP, clinical phrase; LDA, latent Dirichlet allocation; MCC, Mayo Clinical Corpus; MFS, most-frequent-sense; PPR, use of the whole graph but PageRank is run with a non-uniform ('personalized') initial weight vector; SNOMED, Systematized Nomenclature of Medicine; SPR, a subgraph is created based on the context, and traditional PageRank is run.

Topic-modeling results

We use 50-fold cross-validation to maximize the use of labeled examples. Following standard practice, we perform cross-validation on the mapping set, leaving a held-out dataset for testing. Table 2 shows micro- and macro-averages from selected cross-validation runs. We do not report CIs for cross-validation accuracies, since they are only used to select system configurations.

For LDA, the number of senses and the size of the context window were selected over the BOW and bag-of-CPs features, which we refer to as 'base configurations'. We compared the following base configurations: (1) 20 closest words (20w) or six closest words (6w) to the target, excluding the stopwords; (2) 20 closest CPs (20c) or six closest CPs (6c) to the target.

The 20w/20c configurations were chosen for comparison with some of the graph-based methods, as well as with the similar work in the general domain, which used 20-word contexts.[25] The 6w/6c configurations were selected in contrast with the larger window size in order to investigate the impact of a near-minimal context, and are similar to some of the small window sizes used in the general domain.[25] For each context window, the number of latent topics was fixed at the point where the performance plateaued. This number was tuned over all targets, rather than per target. Note that in a practical application, the number of topics can be tuned separately for each target word, likely improving the quality of the resulting clusters. Motivated by the similar topic range over which Brody and Lapata[25] tuned topics for their general domain LDA work, we started at four topics and increased to seven. The 6w and 20w LDA configurations performed comparably, achieving best performance at six topics.

The best LDA configuration was selected by successively adding syntactic and ontological features to the best base configuration. Word-populated syntactic features within three hops are denoted 'Synt'. Syntactic features populated with UMLS semantic types that use just the direct semantic type of the CP are denoted 'UST' ('UMLS semantic type'). Features that instead use the full path to the root of the semantic type hierarchy are denoted 'USTall'. Features populated with UMLS CUI ancestors using k parents are denoted 'UAk' ('UMLS Ancestors k'). Combinations of features are denoted with '+' before each additional set. The best LDA configurations were 6w+6c and 6w +6c+Synt, for six topics. HDP configuration selection was performed over the BOW features and the features that performed best for LDA.

Table 2 MCC cross-validation accuracies

Configuration	Cross-validation macro-average accuracy (%)	Cross-validation micro-average accuracy (%)
MFS	60.1	69.1
LDA, 6w, 6 topics	65.7	73.2
LDA, 20w, 6 topics	65.7	73.2
LDA, 6w+6c+Synt, 6 topics	**66.5**	**73.9**
LDA, 6w+6c, 6 topics	66.0	73.5
LDA, 6w+UST+Synt, 6 topics	65.0	72.7
LDA, 6w+USTall+Synt, 6 topics	61.8	70.3
LDA, 6w+UA2+Synt, 6 topics	60.4	69.2
LDA, 20w+UST+Synt, 6 topics	65.5	73.1
LDA, 20w+USTall+Synt, 6 topics	63.4	71.5
LDA, 20w+UA0+Synt, 6 topics	65.6	73.1
LDA, 20w+UA1+Synt, 6 topics	64.4	72.2
LDA, 20w+20c, 6 topics	65.3	73.0
HDP, 6w+6c+Synt	**70.2**	**76.2**
HDP, 6w+6c	69.7	76.1
HDP, 6w	68.5	75.4
HDP, 20w	65.5	72.8

Best-performing configurations are given in bold.
HDP, hierarchical Dirichlet process; LDA, latent Dirichlet allocation; MCC, Mayo Clinical Corpus; MFS, most-frequent-sense. For explanation of configuration abbreviations see paragraphs 2 and 4 of the 'Topic-modeling results' section.

The failure of ontology-based features UAk to help disambiguation may suggest noisy 'parent' relations. We investigated this by using the paths-to-root that UMLS provides for CUIs in the SNOMED CT vocabulary instead of using parent relations. This leaves the context CPs' CUIs not in SNOMED without any features, but the features that are generated are less noisy. We call USA2 the regeneration of UA2 this way and compare the configurations (1) LDA, 6w+UA2+Synt, six topics and (2) LDA, 6w+USA2+Synt, six topics. The former, with UA2, had 60.4% cross-validation accuracy; the latter, with USA2, had 64.5% cross-validation accuracy. This higher accuracy is still lower than BOW, however (65.7%), so noisy relations must not be the only problem.

We chose to test the best basic configurations (6w) and the best configurations overall (6w+6c+Synt) for both LDA and HDP, shown in table 3. We present both macro- and micro-average accuracies with respect to the targets, as well as 95% CIs for the micro-average. In general, we prefer to evaluate using macro-averages because micro-averages heavily skew the overall performance towards that of 'ms' and 'sob', the targets with 1000 instances in MCC. In the LDA test runs, the extra features showed a gain over just BOW, reflecting cross-validation results. This gain is small but significant ($p=0.0176$). However, HDP test runs showed the opposite trend; this may be partially due to the small number of instances in the test set, as the difference between HDP, 6w+6c+Synt and HDP, 6w is not significant ($p=0.0643$). The difference between the best HDP configuration on the test set, 6w, and its LDA counterpart, 6w with 6 topics, is significant ($p=0.0020$).

Table 4 shows a direct comparison on identical data of these best topic-modeling methods with the graph-based methods investigated earlier. This comparison was performed on our 30% MCC test set limited to the 15 targets on which we evaluated graph-based methods for a total of 444 instances. Again we report macro- and micro-average accuracies and 95% CIs for the latter.

Table 3 MCC test set accuracies

Configuration	Test macro-average accuracy (%)	Test micro-average accuracy (%)	95% CI
MFS	66.7	76.2	74 to 78
LDA, 6w+6c+Synt, 6 topics	**76.9**	**83.4**	**82 to 85**
LDA, 6w, 6 topics	75.0	82.1	80 to 84
HDP, 6w+6c+Synt	76.4	83.0	81 to 85
HDP, 6w	**78.1**	**84.4**	**83 to 86**

Best-performing configurations are given in bold.
HDP, hierarchical Dirichlet process; LDA, latent Dirichlet allocation; MCC, Mayo Clinical Corpus; MFS, most-frequent-sense. For explanation of configuration abbreviations see paragraphs 2 and 4 of the 'Topic-modeling results' section.

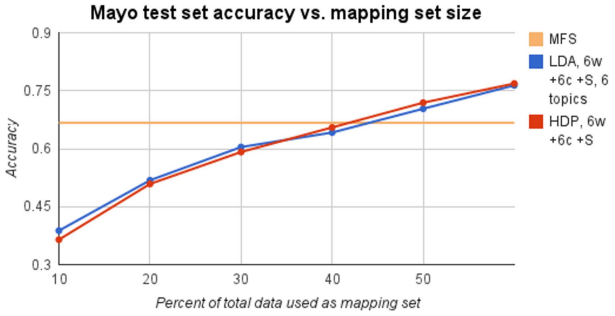

Figure 3 Mayo Clinical Corpus test set accuracy versus mapping set size. HDP, hierarchical Dirichlet process; LDA, latent Dirichlet allocation.

One of the expected effects of the increasing size of the mapping set is that the mapping accuracy would increase until eventually reaching a plateau. Figure 3 shows the MCC test set accuracy as a function of the mapping set size. The accuracy has not plateaued, suggesting that performance may continue to improve with increased mapping set size.

In order to investigate how the topic-modeling approaches compare with the supervised approaches using the same amount of data, we trained a support vector machine (SVM) classifier with BOW features derived from the entire available context using the same 70%/30% split of the MCC data; the mapping sets were used for training and the test sets for evaluation. Since the total number of annotated instances was relatively modest, we used a linear kernel. The resulting macro- and micro-average accuracy was 62.5% and 71.9% across all targets, respectively. The corresponding BOW-only results with six-word context for our topic-modeling approaches on the same split were 75.0%/ 82.1% for LDA and 78.1%/ 84.4% for HDP (Table 3). To the best of our knowledge, the only other supervised learning result available for MCC is due to Savova *et al*.[3] They report 82.6% micro-average accuracy across all targets, with best-performing feature configurations selected separately for each target word. This additional tuning contributes to better performance, but, as they mention, is, in practice, too costly to perform for all targets.[3] In order to evaluate possible effects of the method we used to map the derived topics into senses (cf the section on evaluation metrics), we trained a linear SVM classifier with topic features derived by our best LDA configuration. The resulting macro- and micro-average accuracy across all targets was 75.6% and 82.4%, respectively, which is very close to the corresponding accuracies reported above for the mapping method used in the SemEval WSI task.[16]

Table 4 Comparison of the methods producing the best MCC test set accuracies with graph-based methods on the MCC test set limited to SNOMED CT subset targets

Configuration	Macro-average accuracy (%)	Micro-average accuracy (%)	95% CI
MFS	58.8	59.2	55 to 64
PPR, 20 closest CPs, all relations, initial vector weighted by inverse distance	48.0	48.6	44 to 53
SPR, 20 closest CPs, relations from path-to-hierarchy-root, initial vector weighted by inverse distance	41.9	42.6	38 to 47
Path, CPs in 70 word window, similarity measure wup, uniform CP weighting	44.0	44.4	40 to 49
LDA, 6w+6c+Synt, 6 topics	75.8	76.1	72 to 80
HDP, 6w	76.0	76.4	72 to 80

CP, clinical phrase; HDP, hierarchical Dirichlet process; LDA, latent Dirichlet allocation; MCC, Mayo Clinical Corpus; MFS, most-frequent-sense; SNOMED CT, Systematized Nomenclature of Medicine—Clinical Terms. For explanation of configuration abbreviations see paragraphs 2 and 4 of the 'Topic-modeling results' section.

DISCUSSION

In our results on MCC, of the knowledge-based features, only bag-of-CPs (6c) produced any gain above the BOW baselines. This implies that UMLS only helped disambiguation in identifying and consolidating CPs, and that its graphical properties, which were used in features UST, USTall, and UAk, were unhelpful or harmful. This is perhaps not surprising given the poor performance of our graph-based disambiguation methods, which rely completely on UMLS relations. The fact that even limiting ancestor features to those of CUIs in the SNOMED CT vocabulary's hierarchy tree does not produce a higher average accuracy than BOW suggests that something other than the high fan-out causes problems. It is surprising that the UMLS semantic type hierarchy did not prove helpful, because it is relatively small and its quality is easier to control. Its clustering may be too coarse for use in a WSI task with as fine sense distinctions as are present in MCC (eg, three of the four senses for target 'iv' all relate to slightly different aspects of the same basic meaning). Another factor affecting the poor performance of linguistic features is that the ClearNLP dependency parser used in this work was trained on longer clinical notes and pathology reports, while the training data from the MIMIC II database contains many examples from very abbreviated and shorthand-rich nursing notes.

As always, comparisons between different methods should be taken with a caveat regarding the dataset size. As Banko and Brill pointed out in SemEval-2007,[34] an increase in the amount of training data often trumps the performance increase because of selecting a better algorithm, until eventually a performance plateau is reached, where adding more data no longer helps as much. However, as Peter Norvig argues in his updated version of Banko and Brill's result,[35] a larger volume of data does lead to continued improvement, but on a log scale. And using a different technique (for POS tagging, in Norvig's case) makes less difference than another order of magnitude more data. Which is to say that the argument that it is worth optimizing by choosing the best technique makes sense if you run into limits on availability of data, where you cannot obtain 10× more. This is

definitely the case for WSD data, where sense-tagging of large corpora for every ambiguity is simply not feasible.

Still, the data size effects are definitely relevant for the comparison between different topic-modeling configurations, since the amount of data available for mapping and evaluation in MCC does not allow the systems to reach a plateau in accuracy. This is, perhaps, less relevant for the graph-based approaches, since they require no training data. However, increasing the evaluation set, particularly expanding it to other target words, may well shift the results, since the structure of the UMLS concept graph varies greatly for different targets. Even with the limited amount of evaluation data, though, the difference in performance between the topic-modeling and UMLS graph-based approaches is evident.

The comparison we make between the unsupervised knowledge-based approaches and the topic-modeling approach from the general domain is a comparison between unsupervised methods and semi-supervised methods, as our evaluation of the topic models requires a small amount of labeled data. However, at their core, the topic-modeling algorithms are unsupervised, inducing clusters from the data. The labeled data make it possible to map these clusters, often many-to-one, on to real-world senses, but this may not be necessary in all applications. For example, while a uniform interpretation of a particular natural language query over a set of clinical records may require mapping to a standardized sense inventory, when an ambiguous term is used as a part of a query, other terms from the query may provide sufficient context to retrieve the records with the right sense of the term. Similarly, if WSD is used as an intermediate step in a particular information-extraction task, one may benefit—for example, from using 'disambiguated BOW' features in place of the regular BOW. In these cases, mapping the sense clusters to an existing sense inventory may be irrelevant and unnecessary.

While a full-scale comparison of the topic-modeling approaches with the supervised techniques is outside the scope of this paper, the experiment with an SVM classifier using BOW features does support the notion that such semi-supervised approaches are likely to obtain a higher accuracy with the same amount of data.

CONCLUSIONS

The experiments we present suggest that unsupervised WSI methods using Bayesian topic modeling outperform methods using the UMLS directly. We have also shown that such methods benefit from additional features, particularly syntactic relations and clinically relevant words and phrases ('bag of CPs'), but these techniques would clearly benefit from better integration of the domain knowledge sources. Since UMLS has its weaknesses, future work on features might benefit from using automatic thesaurus construction to more realistically represent relations between CPs, which may then be used to aid disambiguation.

Recent work by Moon *et al*[24] suggests that coarse-grained acronym-related ambiguities may be resolved in a supervised learning framework with even a small number of training examples, provided that a sufficiently large context window (80 closest words) is used. The fact that topic-modeling techniques perform well with a restricted context window (six closest words) suggests that this approach may prove more suitable for clinical applications where little context is available, such as applications supporting natural language queries over text data.

Acknowledgements The authors thank Guergana Savova for assistance with obtaining the sense-annotated data.

Contributors RC is the primary author, responsible for designing, implementing and running the experiments, and producing the first draft of the manuscript. AR is responsible for designing and directing the study, conducting data analyses, and providing guidance throughout the project, as well as providing substantial edits to the manuscript. OU provided input on the methodology and contributed to the preparation of the manuscript. PS offered insights and guidance throughout the project and contributed to the preparation of the manuscript.

Funding This work was supported in part by NIH grant U54-LM008748 from the National Library of Medicine, by contract number 90TR0002 (SHARP—Secondary Use of Clinical Data) from the Office of the National Coordinator (ONC) for Health Information Technology, and by a Siebel Scholar award to Rachel Chasin.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008;35:128–44.
2 Saeed M, Villarroel M, Reisner AT, *et al*. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med* 2011;39:952–60.
3 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.
4 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Suppl 1):D267–70.
5 Aronson AR, Lang FMM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17:229–36.
6 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 Inform Sci Tech* 2006;57:96–113.
7 Humphrey SM, Lu CJ, Rogers WJ, *et al*. Journal descriptor indexing tool for categorizing text according to discipline or semantic type. *AMIA Annu Symp Proc* 2006;2006:960.
8 Agirre E, Soroa A, Stevenson M. Graph-based Word Sense Disambiguation of biomedical documents. *Bioinformatics* 2010;26:2889–96.
9 McInnes BT, Pedersen T, Liu Y, *et al*. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. *AMIA Ann Symp Proc* 2011;2011:895.
10 Jimeno-Yepes A, Aronson A. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics* 2010;11:565.
11 Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009:33–41.
12 Stevenson M, Agirre E, Soroa A. Exploiting domain information for Word Sense Disambiguation of medical documents. *J Am Med Inform Assoc* 2011;19:235–40.
13 Page L, Brin S, Motwani R, *et al*. *The PageRank citation ranking: bringing order to the Web*. Stanford InfoLab, 1999.
14 Savova GK, Pedersen T, Purandare A, *et al*. Resolving ambiguities in biomedical text with unsupervised clustering approaches. Technical report, University of Minnesota Supercomputing Institute Research Report UMSI 2005/80 and CB Number 2005/21. 2005.
15 Savova GK, Therneau T, Chute C. Cluster stopping rules for word sense discrimination. In: *Proceedings of the workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*. Association for Computational Linguistics, 2006:9–16.
16 Manandhar S, Klapaftis IP, Dligach D, *et al*. SemEval-2010 task 14: Word sense induction & disambiguation. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2010:63–8.
17 Pedersen T. Duluth-WSI: SenseClusters Applied to the Sense Induction Task of SemEval-2. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, 2010:363–6.
18 Kern R, Muhr M, Granitzer M. KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, 2010:351–4.
19 Jurgens D, Stevens K. HERMIT: Flexible Clustering for the SemEval-2 WSI Task. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, 2010;359–62.
20 Korkontzelos I, Manandhar S. Uoy: Graphs of unambiguous vertices for word sense induction and disambiguation. In: *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics, 2010:355–8.
21 Pedersen T, Kulkarni A. Automatic cluster stopping with criterion functions and the gap statistic. In: *Proceedings of the Demonstration Session of the Human Language Technology Conference and the Sixth Annual Meeting of the North American*

Chapter of the Association for Computational Linguistics, HLT/NAACL-2006. New York, NY, 2006:276–9.
22 Xu H, Stetson PD, Friedman C. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. *AMIA Annu Symp Proc* 2012;2012:1004–13.
23 Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annual Symposium Proceedings* 2005:589–93.
24 Moon S, Pakhomov S, Melton GB. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. *AMIA Annu Symp Proc* 2012;2012:1310–19.
25 Brody S, Lapata M. Bayesian word sense induction. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009:103–11.
26 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003;3:993–1022.
27 Yao X, Durme BV. Nonparametric Bayesian Word Sense Induction. In: *Graph-based Methods for Natural Language Processing*. The Association for Computer Linguistics, 2011:10–14.
28 Teh YW, Jordan MI, Beal MJ, *et al*. Hierarchical Dirichlet processes. *J Am Stat Assoc* 2006;101:1566–81.
29 Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* 2001:26–33.
30 Choi JD. ClearNLP. https://code:google.com/p/clearnlp/. Google Code, 2013. Computer software.
31 Wu Z, Palmer M. Verbs semantics and lexical selection. In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994:133–8.
32 Phan XH, Nguyen CT. GibbsLDA++: A C/C++ implementation of latent Dirichl*et a*llocation (LDA). 2007.
33 Wang C, Blei DM. *A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process*. ArXiv e-prints, 2012.
34 Agirre E, Soroa A. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007:7–12.
35 Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *IEEE Intell Syst* 2009;24:8–12.

Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods

Rachel Chasin, Anna Rumshisky, Ozlem Uzuner, et al.

J Am Med Inform Assoc published online January 17, 2014
doi: 10.1136/amiajnl-2013-002133