Interactive Medical Word Sense Disambiguation through Informed Learning

Yue Wang[1], Kai Zheng[2], Hua Xu[3], Qiaozhu Mei[1,4]

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA;

[2]Department of Informatics, University of California, Irvine, CA, USA;

[3]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA;

[4]School of Information, University of Michigan, Ann Arbor, MI, USA

Keywords:

Word sense disambiguation

Interactive machine learning

Medical domain knowledge

ABSTRACT

Objective: Medical word sense disambiguation is challenging and often requires significant training data labeled by domain experts. This work aims to develop an interactive learning algorithm that makes efficient use of expert's domain knowledge in building high-quality medical word sense disambiguation (WSD) models with minimal human effort.

Methods: We developed an interactive learning algorithm with experts labeling instances *and features*. An expert can provide supervision in three ways: labeling instances, specifying indicative words of a sense, and highlighting supporting evidence in a labeled instance. The algorithm learns from these labels and iteratively selects the most informative instances to ask for future labels. Our evaluation used three WSD corpora: 198 ambiguous terms in biomedical literature (MSH), 74 ambiguous abbreviations in clinical notes from University of Minnesota (UMN), and 24 ambiguous abbreviations in clinical notes from Vanderbilt University Hospital (VUH). For each ambiguous term and each learning algorithm, a learning curve that plots the accuracy on the test set against the number of labeled instances was generated. The area under the learning curve was used as the primary evaluation metric.

Results: Our interactive learning algorithm significantly outperformed active learning, the previous fastest learning algorithm for medical WSD. Compared to active learning, it achieved 90% accuracy on MSH with 42% less labeling effort, 35% less labeling effort on UMN, and 16% less labeling effort on VUH.

Conclusion: High-quality WSD models can be efficiently trained with minimal supervision by inviting experts to label informative instances and provide domain knowledge through labeling/highlighting contextual features.

1. INTRODUCTION

Medical documents contain many ambiguous terms, the meaning of which can only be determined from the context. For example, the word “ice” may refer to frozen water, methamphetamine (an addictive substance), or caspase-1 (a type of enzyme); and the acronym “PD” may stand for “peritoneal dialysis” (a treatment for kidney failure), “posterior descending” (a coronary artery), or “police department”. Assigning the appropriate meaning (a.k.a. ‘sense’) to an ambiguous word based on the context is referred to as the process of word sense disambiguation (WSD)[1-2]. WSD is a critical step for many medical natural language processing (NLP) applications, such as text indexing and categorization, named entity extraction, and computer-assisted chart review.

The research community has proposed and evaluated many WSD methods in the past, including supervised learning[3-5], semi-supervised learning[6-8], and knowledge-driven[9-10] approaches. Collectively, these studies have shown that a substantial volume of high-quality training data annotated by human experts is required for existing WSD models to achieve desirable performance. However, annotating training data is a labor-intensive process, and the quality may deteriorate as the volume required to be annotated increases[11]. This is particularly true for medical WSD as assigning correct sense for ambiguous medical terms requires great attention and highly specialized domain knowledge.

To address this issue, the machine learning community has been exploring approaches that involve human experts just-in-time during a machine learning process, in contrast to conventional approaches wherein human experts are only involved in creating static annotated training or evaluation datasets. Such approaches are generally referred to as ‘active’ learning. An active learning approach[12] prioritizes instances to be labeled and presents to human experts the most informative ones that would help the algorithm achieve desirable performance with fewer iterations. This family of learning methods has shown far superior performance over that of random sampling in medical WSD tasks[13].

In our previous work[14], we described ReQ-ReC expert, a step further by incorporating an information retrieval component in active learning that allows human experts to identify and label typical instances using their domain knowledge through keyword search. It demonstrated better performance than active learning in medical WSD tasks. However, even though experts are brought into the loop, existing interactive learning approaches still suffer from the "cold start" problem. That is, without any prior knowledge about a new WSD task, an algorithm based on artificial intelligence (i.e., a statistical WSD classifier) needs a large amount of training data to reach a reasonable accuracy. In contrast, *well-trained* human experts do not have the cold start problem because they come to a WSD task with established domain knowledge, which helps them directly determine the correct sense of an ambiguous word.

In this paper, we describe a novel interactive learning algorithm that is capable of directly acquiring domain knowledge from human experts by allowing them to articulate the evidence that leads to their sense tagging decisions (e.g., the presence of indicative words in the context that suggest the sense of the word). This knowledge is then applied in subsequent learning processes to help the algorithm achieve desirable performance with fewer iterations, thus solving the cold start problem. That is, besides labeling instances, the expert can provide domain knowledge by two means: (1) to specify informative words of a sense, and (2) to highlight evidence words in labeled instances. These interaction modes enable experts to directly express their prior knowledge and thought process when they perform WSD, without adding much burden. The two channels complement each other: it is sometimes hard to specify strong informative words *a priori*, but easier to highlight these words *in situ*. The statistical classifier can learn from both labeled instances and informative words (i.e. labeled features), and query new labels using active learning.

Simulated experiments on three WSD corpora show that expert's domain knowledge gives the model a 'warm start' at the beginning stage, significantly accelerating the learning process. On one biomedical literature corpus and two clinical notes corpora, the proposed algorithm makes better use of human experts in training WSD models than all existing approaches, achieving the state-of-the-art performance with least effort.

2. METHODS

2.1 Instance Labeling vs. Feature Determination

Below, we use an example to illustrate how the interactive learning algorithm works. Suppose the word “cold” (or its spelling variants, e.g., “COLD”) is mentioned across a set of medical documents. Depending on the context, it could mean “chronic obstructive lung disease,” “common colds,” or “low temperature.” The task of WSD is to determine the correct sense of each appearance of this word (i.e., each *instance* of the word).

A human expert performing this task may apply a number of rules based on her or his domain knowledge. For example, she or he may know that when all letters of the word are spelled in capital case, i.e., “COLD,” it is more likely the acronym of “chronic obstructive lung disease” than any other possible senses. This judgment could be further strengthened when there are indicative words (or phrases) such as “chronic,” “obstructive,” or “lung” in the adjacent text. Likewise, if the word is not spelled in all capitals, and is accompanied by words such as “common,” “cough,” and “sneeze,” it likely means “common cold.” For certain senses, contextual cues may appear in other forms rather than indicative words. For example, a numeric value followed by a unit of temperature (e.g. “5 degrees C”) may give out that the word “cold” in the current context likely refers to “low temperature,” instead of a medical condition.

Unfortunately, such domain knowledge is not leveraged by conventional supervised learning approaches, which only ask human experts to label the sense of the instances of an ambiguous word, rather than capture *how* human experts make such judgments. In other words, conventional approaches only try to ‘infer’ human wisdom from annotated results, instead of acquiring it directly—even if such wisdom is readily available and can be formalistically expressed. The interactive learning algorithm described in this paper addresses this limitation by allowing human experts to create *labeled features* in addition to labeling instances.

A *labeled instance* for an ambiguous word is a [*context*, *sense*] pair, following the conventional definition in supervised learning. For example, a labeled instance of the word “cold” can be:

[“The patient developed ***cold*** and experienced cough and running nose.”, common cold] .

A *labeled feature* for an ambiguous word is a [*feature*, *sense*] pair, where the *feature* is a textual pattern (a word, a phrase, a skip *n*-gram, or a regular expression in general). The pair encodes the (most likely) *sense* of the ambiguous word if the *feature* appears in its context. For example, human experts can express domain knowledge of the sense of “cold” by creating the following labeled features:

[“COLD” : All cap,	chronic obstructive lung disease]
[“chronic” : Non all-cap,	chronic obstructive lung disease]
[“obstructive” : Non all-cap,	chronic obstructive lung disease]
[“lung” : Non all-cap,	chronic obstructive lung disease]
[“common” : Non all-cap,	common cold]
[“cough” : Non all-cap,	common cold]
[“sneeze” : Non all-cap,	common cold]

...

Human experts can also express domain knowledge by highlighting a contextual cue after labeling an instance of “cold”, as in

[“The tissue was exposed to a ***cold*** environment (5 degrees C).”, low temperature].

The highlighted text snippet essentially creates another labeled feature for “cold”:

[“<digit> degrees C”,   low temperature] .

A labeled feature encodes certain domain knowledge that human experts use to solve a WSD task, which can be directly applied to train machine-learning models. As a result, it improves WSD performance and, at the same time, reduces the amount of manual effort required to create a large quantity of labeled instances as training data.

2.2 Overall Workflow

The interactive learning algorithm consists of several distinct components; illustrated in Figure 1.

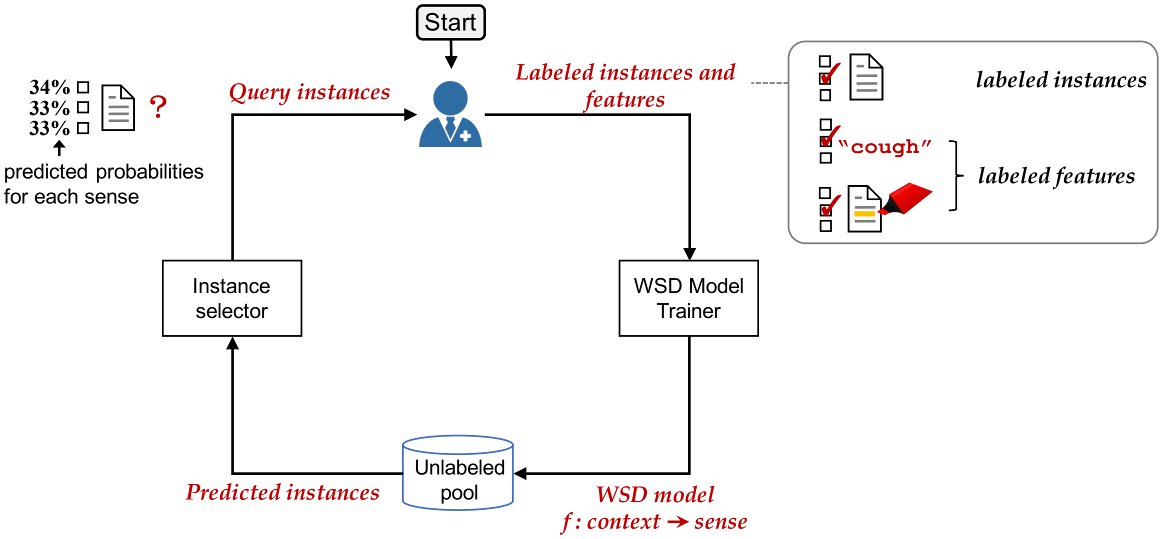

Figure 1. Interactive learning with labeled instances and features

When the human expert can come up with good features for each sense of an ambiguous word, the algorithm can directly use them to train an initial WSD classifier. When such domain knowledge is not available, we assume that the human expert can identify at least one instance for each sense. She or he can then label the instance and highlight contextual cues in that instance. This kicks off the interactive learning process.

The algorithm contains an *instance selector* that determines how to best select instances from an unlabeled pool to present to the human expert. Then, the human expert labels the sense of the instance, followed by potentially suggesting features that were used as the "rationale" for the labeling decision (i.e. feature labeling). Next, the algorithm uses both labeled instances and labeled features to retrain the WSD classifier, then begins another iteration by selecting additional instances for manual labeling till satisfactory WSD result is achieved. This process is described in more detail in the next few sections.

2.3 WSD Model Training

The algorithm of training and retraining a WSD model consists of two stages: feature representation and parameter estimation.

2.3.1 Dynamic Feature Representation

In conventional supervised learning, a model uses a fixed set of features throughout the training process. For text classification, this feature set is often all of the words in the corpus. In our interactive learning algorithm, labeled features may contain arbitrary textual patterns that are difficult to know ahead of time. Rather than trying to include all possible features from the beginning as conventional machine-learning methods do, we use a dynamic feature representation by starting with a set of *base* features and gradually expanding it as new features emerge. This method helps to prevent severe overfitting when the size of the feature set is large.

We use presence/absence of unigrams as the base features to represent an instance: $x^{base} \in \mathbb{R}^V$, where V is the number of distinct unigrams. A labeled feature defines a real-valued function $\phi(\cdot)$ of an instance, such as “1 if the instance contains ‘COLD’ in all caps; 0 otherwise”. Suppose we have *m* labeled features at iteration *t*, then an instance is represented by a (*V*+*m*)-dimension vector $x = [x^{base}, \phi^{(1)}, \cdots \phi^{(m)}]$.

2.3.2 Parameter Estimation

We use logistic regression with linear kernel as the WSD classifier. If an ambiguous word has two senses, we build a binary classifier, otherwise a softmax multiclass classifier. Logistic regression classifiers output probability predictions in $[0,1]$, which are then used by the active learning algorithm.

Below, we describe the algorithm for training the logistic regression model. Suppose at a certain iteration, we have *l* labeled instances $\{(x^{(i)}, y^{(i)})\}_{i=1}^{l}$, and m labeled features $\{(\phi^{(j)}, y^{(j)})\}_{j=1}^{m}$. For an ambiguous word with k senses, $y^{(i)}$ or $y^{(j)}$ is a one-hot *k*-dimensional vector that encodes the assigned sense. We train a logistic regression model $p(y|x;w)$ by minimizing the following loss function (w denotes the parameters of the model):

$$J(w) = \sum_{i=1}^{l} \sum_{c=1}^{k} -y_c^{(i)} \log p\left(y_c \middle| x^{(i)}; w\right) + \lambda_1 \sum_{j=1}^{m} \sum_{c=1}^{k} -\tilde{y}_c^{(j)} \log p\left(y_c \middle| \phi^{(j)}; w\right) + \frac{\lambda_2}{2} \|w\|_2^2 \quad (1)$$

$p(y_c|\phi^{(j)};w)$ is the expectation for any instance containing feature $\phi^{(j)}$ to have sense *c*. Let S_j be the set of instances (both labeled and unlabeled) with non-zero feature values for $\phi^{(j)}$, then

$$p\left(y_c \middle| \phi^{(j)}; w\right) = \frac{\sum_{i \in S_j} p(y_c|x^{(i)}; w)}{|S_j|}.$$

$\tilde{y}_c^{(j)} = (y_c + \varepsilon)/(1 + k\varepsilon)$ is the smooth version of feature label distribution, because unlike labeled instances, labeled features should be interpreted as preferences rather than as absolute assignments. $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are trade-off weights for different loss terms. In this paper, we set $\varepsilon = 0.1, \lambda_1 = \lambda_2 = 1$.

In the loss function (1), the first term is the cross-entropy loss on labeled instances; the second term is the cross-entropy loss on labeled features; and the third term is a regularization term of parameter *w*. If the loss function only consists of the first and the third term, then it reduces to the loss function of a traditional softmax logistic regression classifier. The second term expresses a preference on the expected behavior of the WSD classifier, i.e., the presence of a feature strongly suggests a label (i.e., the most probable sense). This is so called generalized expectation criterion[15]. Because of the second term, (1) is a nonconvex function. We use gradient descent to find a local minimum for the model parameter *w*. In practice, we find the local minimum yields a sufficiently performing classification model.

2.4 Instance Selection

The proposed algorithm kicks off the first iteration by a labeled feature for each sense. Once the WSD classifier $p(y|x;w)$ is trained, active learning can be applied to select a small set of unlabeled instances to present to human experts for labeling. Specifically, we use minimum margin-based active learning as the instance selection algorithm which has shown superior performance in classification settings[12,14]. It selects the unlabeled instance x that satisfies the smallest $Q(x) = p(y_1|x;\theta) - p(y_2|x;\theta)$, where y_1 and y_2 are the most and second most probable senses. Intuitively, the classifier cannot determine whether y_1 or y_2 is the correct sense, therefore it needs to solicit input from human experts.

2.5 Evaluation Method

2.5.1 Evaluation Corpora

In this study, we used three established medical corpora to evaluate the performance of the interactive learning algorithm.

The MSH corpus contains a set of MEDLINE abstracts automatically annotated using MeSH indexing terms[16]. Similar to how it was handled in previous work[13-14], for this corpus, we only included ambiguous words that have at least 100 instances, providing adequate data for training and evaluation. This gave us 198 ambiguous words, including 102 abbreviations, 86 non-abbreviated words, and 10 abbreviation-word combinations.

The UMN corpus contains 74 ambiguous abbreviations from a total of 604,944 clinical notes created at the Fairview Health Services affiliated with the University of Minnesota; each abbreviation has 500 randomly sampled instances[17]. Each instance is a paragraph in which the abbreviation appeared. 4 abbreviations have a general English sense (*FISH*, *IT*, *OR*, *US*).

The VUH corpus contains ambiguous abbreviations from the admission notes created at the Vanderbilt University Hospital[18]. Similar to the MSH corpus, we only retained 24 abbreviations that have more than 100 instances. Each instance is a sentence in which the abbreviation appeared. One abbreviation is a loanword in English (*AD* as in “ad lib”).

The summary statistics of these three evaluation corpora is shown in Table 1 (more details can be found in online appendix tables A1-A3). The MSH corpus has the richest context in an instance (i.e., highest average number of tokens per instance), and the least skewed distribution of senses (i.e., lowest proportion of dominating majority senses). Because the main objective of this study was to evaluate the performance of the interactive learning algorithm in comparison with other machine-learning algorithms, we did not further tune the context window size for each corpus. The three corpora share 3 abbreviations (*SS*, *CA*, *RA*). MSH and UMN share another 6 abbreviations. UMN and VUH share another 5 abbreviations. The same abbreviation may have different senses in different corpora.

Table 1. Summary statistics of three evaluation corpora.

	Corpus size	**Average number of instances per word**	**Average number of senses per word**	**Average number of tokens per instance**	**Average percentage of majority sense**
MSH	198	190	2.1	202.84	54.2%
UMN	74	500	5.5	60.59	73.4%
VUH	24	194	4.3	18.73	78.3%

2.5.2 Baseline Methods

To comparatively evaluate the performance of the interactive learning algorithm, we included three other machine-learning algorithms in the analysis. As shown in Table 2, these algorithms vary mainly based on how labeled instances or features are obtained from human experts.

Table 2. Description of baseline methods.

Random sampling	**Active learning**	**ReQ-ReC expert**	**Informed learning**
The algorithm selects the next instance at random from the unlabeled pool.	The algorithm selects the next instance using the minimum margin criterion[12-13].	The algorithm extends active learning by inviting human experts to search for typical instances for each sense using keywords[14].	The proposed interactive learning algorithm.

Start with one labeled instance for each sense.	Start with one labeled instance for each sense.	Start with one labeled feature for each sense.	Start with one labeled feature (or one labeled instance with a highlighted feature) for each sense.
Later iterations use random sampling to obtain instance labels.	Later iterations use minimum margin to obtain instance labels.	Later iterations use minimum margin to obtain instance labels.	Later iterations use minimum margin to obtain instance labels.

2.5.3 Simulated Human Expert Input

To derive evaluation metrics, we simulated human expert input using labeled data from each corpus, which is a method commonly used to evaluate active learning algorithms[12]. This method reduces potential influences that may be introduced due to performance variation by human experts. More specifically:

(1) **Labeling instances**: We used the validated labels in these evaluation corpora as the oracle of instance labels.

(2) **Labeling features**: To implement simulated human expert input (i.e. the 'oracle') that ***provides*** labeled features, we computed information gain for each unigram feature using the entire labeled corpus[19], and selected the most informative features as oracle features. A feature is associated with a sense when the feature co-occurs most frequently with the sense. To make it more realistic, we simulated the oracle that knows the q-th best feature among all unigram features, where q = 1, 5, 10. This oracle was also used in the "ReQ-ReC expert" algorithm when composing the first search query. The labeled features generated in this way were mostly the words in the definition of each sense.

Since in reality, a human expert is unlikely able to come up with all features achieving the highest information gain, we also implemented a weaker, supplementary oracle that better resembles true human performance in realistic WSD tasks. It simulates the action of the expert ***highlighting*** a feature in a labeled instance while she or he is doing the annotation. In the first iteration, a random instance in each sense was given to the oracle. It identified the most informative *n*-gram (*n*=1,2,3) feature in that instance. We used *n*-grams instead of unigrams to allow the oracle to highlight consecutive words in a sentence. To make the oracle more realistic, we simulated the oracle that knows the *q*-th best *n*-gram feature in that instance, where *q* = 1, 2, 3.

2.5.4 Evaluation Metrics

We used learning curves to evaluate the cost-benefit performance of different learning algorithms. A learning curve plots the learning performance against the effort required in training the algorithm. In the context of this paper, learning performance is measured by classification accuracy on a test corpus; and effort is measured by the number of instances that need to be labeled by human experts. For each ambiguous word, we split its instances into an unlabeled set and a test set. When a learning algorithm is executed over the unlabeled set, a label is revealed only if the learning algorithm asks for it. With more and more labels becoming available, the WSD model is continuously updated and its accuracy continuously evaluated, producing a learning curve.

To reduce variation of the curve due to differences between the unlabeled set and the test set, we ran a 10-fold cross validation: 9 folds of the data are used as the unlabeled set and one fold used as the test set. The learning curve of the algorithm on a particular ambiguous word is produced by taking the average of the 10 curves. The overall aggregated learning curve of the algorithm is obtained by taking the average of all curves on all ambiguous words in an evaluation corpus.

In reality, human experts are unlikely to provide an inclusive set of features with the highest information gain prior to the annotation process. On the other hand, a well-trained human annotator should be able to identify the best (or one of the best) features after seeing and labeling an instance. Therefore, we hypothesize that the true performance of a human expert will be between the oracle that provides the best feature (best-case scenario) and the oracle that highlights the 3rd best feature in a labeled instance (worst-case scenario). We average the learning curves of the best- and the worst-case scenarios to generate the learning curve of "informed learning".

To summarize the performance of different learning algorithms using a composite score, we also generated a global Area under Learning Curve (ALC) for each algorithm on each corpus. This method was introduced in the 2010 Active Learning Challenge[20]. The global ALC score was normalized by the area under the best achievable learning curve (constant 1.0 accuracy over all points).

To test the significance of performance difference between the algorithms in terms of average ALC scores, we used Wilcoxon signed rank test[21], a non-parametric test for paired examples. We set the type I error control at $\alpha = 0.01$.

3. RESULTS

The aggregated learning curves obtained by applying each of the learning algorithms on the evaluation corpora, including drill-down analyses of imperfect feature labeling and highlighting oracles, are exhibited in Figures 2–4.

The learning curves of informed learning algorithm demonstrated a "warm start" substantially better than the other algorithms evaluated. This is as a result of applying directly acquired domain knowledge from human experts at the beginning of the learning process. The warm start not only helps to achieve desired performance faster with fewer instance labels, but also makes the proposed algorithm (potentially) less susceptible to highly skewed sense distribution. As shown by the curves on the two clinical WSD corpora, UMN and VUH. To reach 90% accuracy, informed learning saved 42% instance labels compared to active learning on the MSH corpus (15 vs. 26), 35% instance labels on the UMN corpus (15 vs. 23), and

16% instance labels on the VUH corpus (26 vs. 31).

The ALC scores for each corpus and each learning algorithm, as well as the results of statistical significance tests, are reported in Table 3. On all three corpora, Wilcoxon signed rank test showed that the ALC scores of informed learning were statistically significantly better than margin-based active learning. On two corpora (MSH and UMN), the ALC scores of informed learning were statistically significantly better than ReQ-ReC expert, the previous state of the art. These significance results hold even when the feature oracles were imperfect, demonstrating that the proposed algorithm was applicable in a broad range of conditions.

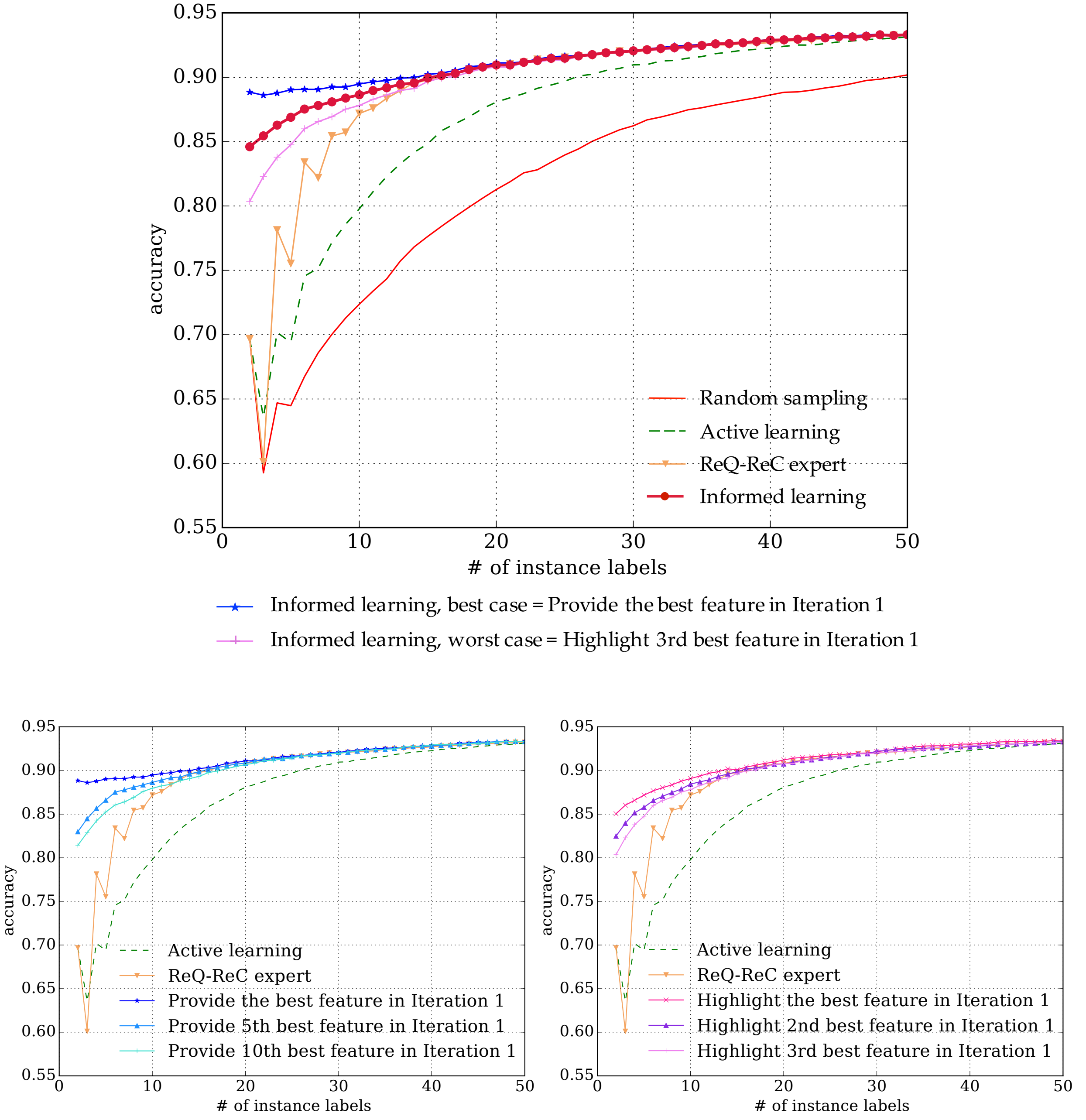

Figure 2. Aggregated learning curves of 198 ambiguous words in the MSH corpus. **Top:** interactive learning algorithms in comparison, including the best- and worst-case scenarios of "informed learning". To achieve 90% accuracy, "random sampling" required 49 instance labels, and "active learning" required 26 instance labels. "ReQ-ReC expert" used labeled features as instance search queries and required 17 instance labels to achieve 90% accuracy. "Informed learning" directly learned from feature labels and

only required 15 instance labels to achieve 90% accuracy. **Lower left (right):** drill-down analysis of informed learning using imperfect feature labeling (highlighting) oracles, respectively. Even using imperfect feature labeling oracles, variants of “informed learning” still significantly outperformed both “active learning” and “ReQ-ReC expert”, according to Wilcoxon signed rank test (see Table 3).

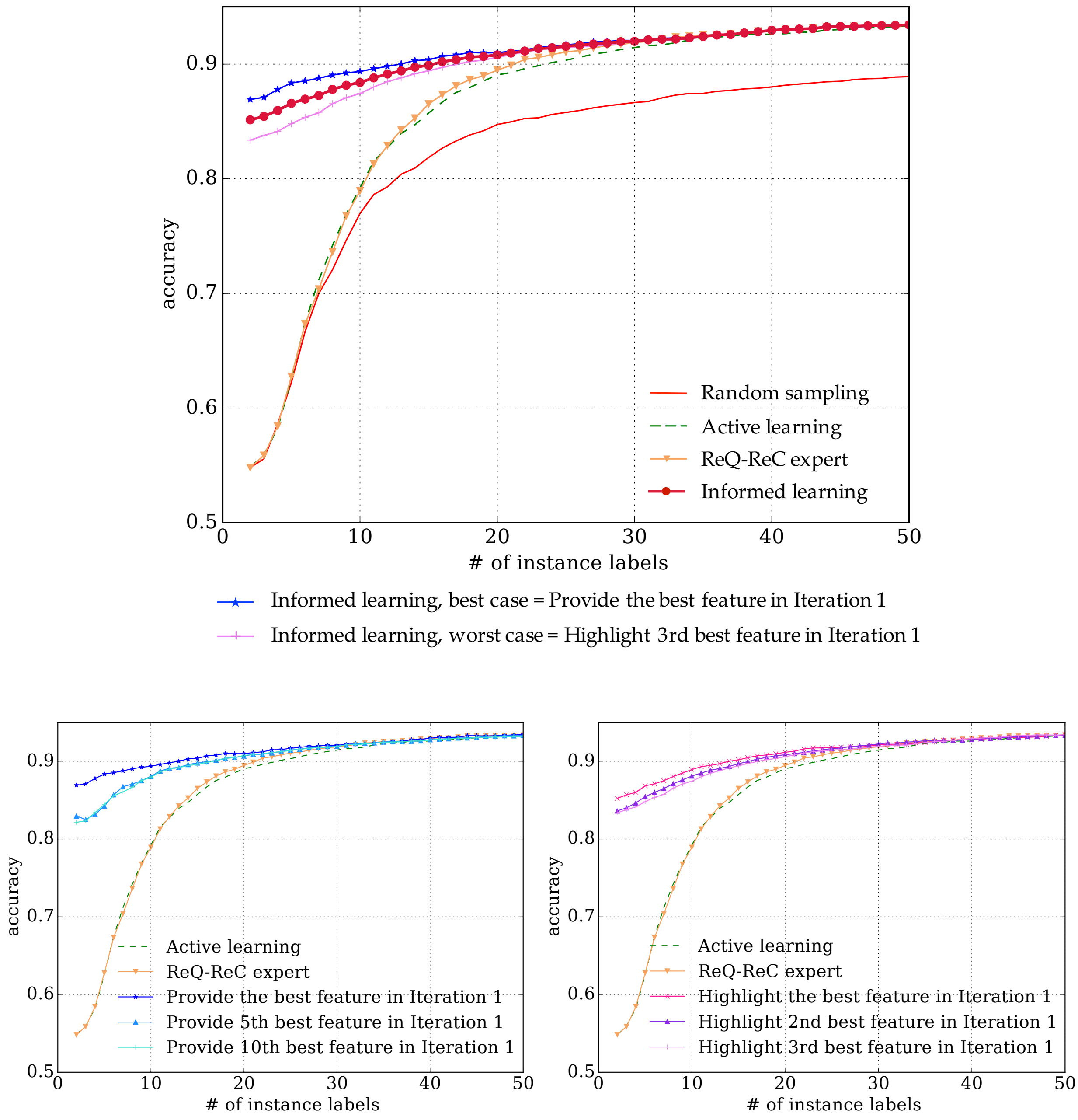

Figure 3. Aggregated learning curves of 74 ambiguous words in the UMN corpus. **Top:** interactive learning algorithms in comparison, including the best- and worst-case scenarios of "informed learning". To achieve 90% accuracy, "random sampling" required more than 50 instance labels, "active learning" required 23 instance labels, and "ReQ-ReC expert" required 21 instance labels. "Informed learning" required only 15 instance labels. **Lower left (right):** drill-down analysis of informed learning of imperfect feature labeling (highlighting) oracles, respectively. Even using imperfect feature oracles,

variants of “informed learning” still significantly outperformed both “active learning” and “ReQ-ReC expert”, according to Wilcoxon signed rank test (see Table 3).

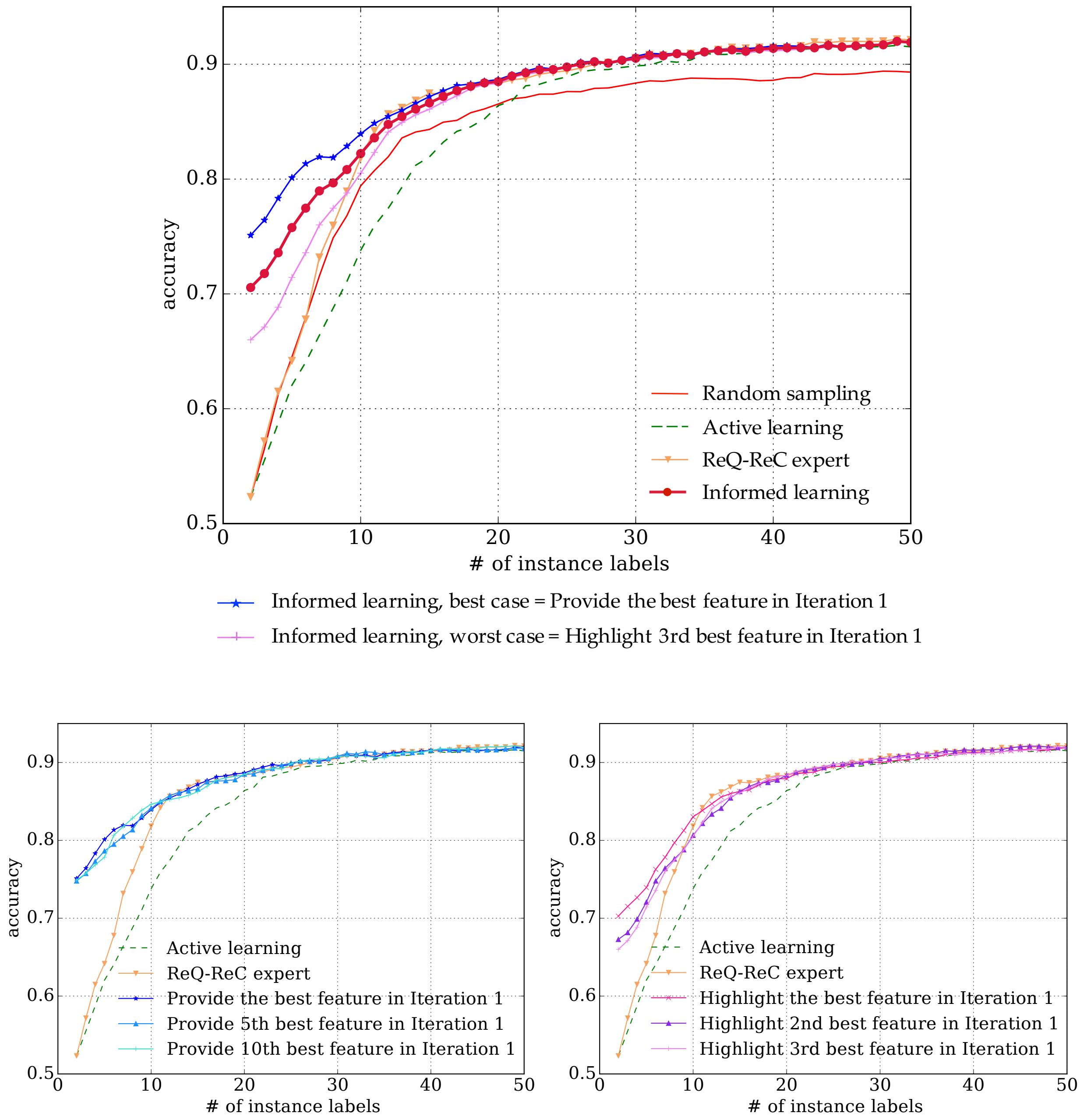

Figure 4. Aggregated learning curves of 24 ambiguous words in the VUH corpus. **Top**: interactive learning algorithms in comparison, including the best- and worst-case scenarios of "informed learning". To achieve 90% accuracy, "random sampling" required more than 50 instance labels, "active learning" required 31 instance labels, "ReQ-ReC expert" and "Informed learning" required 26 labels. **Lower left (right)**: drill-down analysis of learning curves of imperfect feature labeling (highlighting) oracles,

respectively. Even using imperfect feature oracles, variants of "informed learning" still significantly outperformed "active learning", according to Wilcoxon signed rank test (see Table 3).

Table 3. Area under learning curve (ALC) scores of evaluated interactive learning algorithms. The bottom two sections are variants of "Informed learning" with different feature labeling (highlighting) oracles. '*' means the score is significant compared to "Active learning" at level $\alpha = 0.01$. '†' means the score is significant compared to "ReC-ReQ expert" at level $\alpha = 0.01$.

	MSH	UMN	VUH
Random sampling	0.8159	0.8146	0.8311
Active learning	0.8676	0.8522	0.8309
ReQ-ReC expert	0.8928	0.8550	0.8524
Informed learning	$0.9094^{*\dagger}$	$0.9074^{*\dagger}$	0.8706^{*}
Provide the best feature in Iteration 1	$0.9141^{*\dagger}$	$0.9122^{*\dagger}$	0.8792^{*}
Provide 5th best feature in Iteration 1	$0.9087^{*\dagger}$	$0.9038^{*\dagger}$	0.8773^{*}
Provide 10th best feature in Iteration 1	$0.9052^{*\dagger}$	$0.9029^{*\dagger}$	0.8777^{*}
Highlight the best feature in Iteration 1	$0.9119^{*\dagger}$	$0.9091^{*\dagger}$	0.8675^{*}
Highlight 2nd best feature in Iteration 1	$0.9072^{*\dagger}$	$0.9035^{*\dagger}$	0.8639^{*}
Highlight 3rd best feature in Iteration 1	$0.9047^{*\dagger}$	$0.9004^{*\dagger}$	0.8620^{*}

4. DISCUSSION

Warm-start effect. The informed learning algorithm is perfectly positioned to address the "cold start" problem. Active learning works best when the model has a reasonably good "understanding" of the problem space so that the selected instances are the most informative. At the beginning, the model trained on very few labeled instances can perform poorly and waste data selection. In informed learning, human experts can start the learning process by specifying an informative keyword of a sense, which essentially provides weak labels to many instances containing that keyword, resulting in a "warm start". It significantly reduces total number of instance labels to reach high accuracy.

Error analysis. In Table 4, we break down the performance of each algorithm on different subsets of words in three corpora. In the MSH corpus, as abbreviations often co-occur with its full forms, they were easier to disambiguate than non-abbreviated words. The abbreviations in UMN and VUH were harder to disambiguate than those in MSH, because the unbalanced sense distribution presented a challenge to machine learning models.

We studied the cases where Informed Learning (IL) underperformed Active Learning (AL) or ReQ-ReC expert (RR). The main reason was that the simulated feature oracle sometimes provided low-quality labeled features. In fact, words with high information gain could be rare words, not generalizing to many examples; they could also be common words (e.g., “that”, “of”), which happened to appear more frequently in one sense than others but were too noisy to be useful in classification. IL works well when a labeled feature is representative of and specific to a sense. We hypothesize that real human experts are more capable of providing such high-quality features than simulated experts.

AL and RR start learning with equal number of instances in each sense, i.e. assuming a uniform prior distribution over senses. As for IL, initial labeled features induce a sense distribution through feature popularity (a frequent feature indicates a major sense), naturally giving rise to a skewed sense distribution. When the true sense distribution is indeed uniform (MSH), AL and RR may have an advantage over IL. However, when the true sense distribution is skewed (UMN and VUH), AL and RR may suffer as they need more instance labels to correct their uniform prior assumption.

Table 4. Average ALC scores of evaluated interactive learning algorithms across different subsets of ambiguous words.

Subsets of ambiguous words in each corpus		Average ALC score				ALC advantage (%)	
		Random sampling	Active learning	ReQ-ReC expert	Informed learning	Informed over Active	Informed over ReQ-ReC
MSH	102 abbreviations	0.8617	0.9189	0.9349	0.9548	101/102 (99%)	98/102 (96%)
	10 abbreviation-word combinations	0.8265	0.8623	0.8922	0.9150	10/10 (100%)	10/10 (100%)
	86 non-abbreviated words	0.7603	0.8074	0.8430	0.8549	86/86 (100%)	66/86 (77%)
UMN	70 abbreviations	0.8145	0.8520	0.8545	0.9076	70/70 (100%)	70/70 (100%)
	4 abbreviation-word combinations	0.8176	0.8540	0.8635	0.9048	4/4 (100%)	4/4 (100%)
VUH	23 abbreviations	0.8332	0.8343	0.8552	0.8710	21/23 (91%)	18/23 (78%)
	1 abbreviation-word combination	0.7820	0.7535	0.7877	0.8490	1/1 (100%)	1/1 (100%)

In this study, we set 90% accuracy as the target and measured the number of instances required for achieving that performance. In secondary analysis of EHRs data for clinical research, NLP systems with over 90% accuracy are often viewed as reasonable[22-24] and have been widely used. However, for NLP systems that will be used for clinical practice (e.g., clinical decision support systems), higher performance would be required. Therefore, the target performance is dependent on specific tasks. In the future, we will further investigate our approaches when required performance changes.

5. CONCLUSION

This paper introduces a novel interactive machine learning algorithm that can learn from domain knowledge to rapidly build statistical classifiers for medical WSD. Human experts can express domain knowledge by either prescribing informative words for a sense, or highlighting evidence words when labeling an instance. In addition, active learning technique is employed to query instance labels. Experiments using three biomedical WSD corpora showed that the algorithm delivered significantly better performance than strong baseline methods. In the future, we will conduct evaluation studies to assess the performance of the algorithm using real-world scenarios with real human experts.

Funding

This work was supported by the National Science Foundation under grant number 1054199; and the National Library of Medicine under grant number 2R01LM010681-05.

Competing Interests

None.

Contributorship

YW preprocessed the data, designed and implemented the interactive learning algorithms, conducted experiments and statistical significance tests, and drafted and revised the manuscript. KZ revised the experimental design, interpreted the results, and extensively revised the manuscript. HX conceived the research project, provided the data, and extensively revised the manuscript QM conceived the research project, designed the algorithmic framework and evaluation methodology, and extensively revised the manuscript.

REFERENCES

1. Ide N, Véronis J. Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics. 1998;24(1):2–40.
2. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. Journal of Computational Biology. 2005;12(5):554–565.
3. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. Journal of the American Medical Informatics Association. 2004;11(4):320–331.
4. Xu H, Markatou M, Dimova R, et al. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics. 2006;7(1):334.
5. Wu Y, Xu J, Zhang Y, et al. Clinical abbreviation disambiguation using neural word embeddings. Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP). 2015:171-176.
6. Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. Journal of Biomedical Informatics. 2001;34(4):249–261.
7. Xu H, Stetson PD, Friedman C. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. AMIA Annual Symposium Proceedings. vol. 2012. AMIA 2012:1004-13.
8. Finley GP, Pakhomov SV, McEwan R, et al. Towards Comprehensive Clinical Abbreviation Disambiguation Using Machine-Labeled Training Data. In: AMIA Annual Symposium Proceedings. vol. 2016. AMIA 2016:560-569.
9. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. Journal of the American Medical Informatics Association. 2002;9(6):621–636.

10. Yu H, Kim W, Hatzivassiloglou V, et al. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. Journal of Biomedical Informatics. 2007;40(2):150–159.

11. Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. O'Reilly Media, Inc.; 2012.

12. Settles B. Active learning literature survey. University of Wisconsin, Madison. 2010;52(55-66):11.

13. Chen Y, Cao H, Mei Q, et al. Applying active learning to supervised word sense disambiguation in MEDLINE. Journal of the American Medical Informatics Association. 2013;20(5):1001–1006.

14. Wang Y, Zheng K, Xu H, et al. Clinical word sense disambiguation with interactive search and classification. AMIA Annual Symposium Proceedings. vol. 2016. AMIA 2016:2062–2071.

15. Druck G, Mann G, McCallum A. Learning from labeled features using generalized expectation criteria. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2008:595–602.

16. Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics. 2011 Jun 2;12:223.

17. Moon S, Pakhomov S, Liu N, et al. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. Journal of the American Medical Informatics Association. 2014 Mar;21(2):299–307.

18. Wu Y, Denny J, Rosenbloom ST, et al. A Prototype Application for Real-time Recognition and Disambiguation of Clinical Abbreviations. In: Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics. New York, NY, USA: ACM 2013:7–8.

19. Yang Y, Pedersen J O. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning. ACM 1997:412–420.

20. Guyon I, Cawley G, Dror G, et al. Results of the Active Learning Challenge. JMLR: Workshop and Conference Proceedings 2011;16:19–45.

21. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945 Dec 1;1(6):80-3.
22. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association. 2010 Jan 1;17(1):19-24.
23. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Jun 16;18(5):552-6.
24. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association. 2013 Apr 5;20(5):806-13.

Interactive Medical Word Sense Disambiguation through Informed Learning

Yue Wang, Kai Zheng, Hua Xu, Qiaozhu Mei

Appendix

Table A1. Interactive learning results for 198 ambiguous words in the MSH corpus.
Notes:

- Type "A" represents an Abbreviation; type "T" represents a Term; type "AT" represents an Abbreviation and Term.
- #S: number of senses.
- #inst: number of instances.
- RS: Random Sampling.
- AL: Active Learning.
- RR: ReQ-ReC expert.
- IL: Informed Learning.
- "IL > AL": the ALC score of Informed Learning is greater than that of Active Learning. Equals 1 if true; 0 otherwise.
- "IL > RR": the ALC score of Informed Learning is greater than that of ReQ-ReC expert. Equals 1 if true; 0 otherwise.

ID	Word	Type	#S	#inst	#inst in top 5 senses					major sense ratio	ALC scores				IL > AL	IL > RR
					S1	S2	S3	S4	S5		RS	AL	RR	IL		
1	AA	A	2	198	99	99	0	0	0	0.5000	0.8965	0.9385	0.9619	0.9784	1	1
2	ADA	A	2	198	99	99	0	0	0	0.5000	0.8335	0.8883	0.9416	0.9347	1	0
3	ADH	A	2	198	99	99	0	0	0	0.5000	0.9002	0.9512	0.9649	0.9771	1	1
4	ADP	A	2	149	99	50	0	0	0	0.6644	0.8971	0.9166	0.8689	0.9102	0	1
5	Adrenal	T	2	198	99	99	0	0	0	0.5000	0.6467	0.7324	0.7653	0.7608	1	0
6	Ala	A	3	297	99	99	99	0	0	0.3333	0.7698	0.8337	0.8812	0.9149	1	1
7	ALS	A	2	198	99	99	0	0	0	0.5000	0.9152	0.9559	0.9711	0.9785	1	1
8	ANA	A	2	198	99	99	0	0	0	0.5000	0.8769	0.8982	0.9325	0.9430	1	1
9	Arteriovenous Anastomoses	T	2	129	99	30	0	0	0	0.7674	0.8267	0.8724	0.8881	0.9203	1	1
10	Astragalus	T	2	198	99	99	0	0	0	0.5000	0.9259	0.9464	0.9441	0.9606	1	1
11	B-Cell Leukemia	AT	2	158	92	66	0	0	0	0.5823	0.6888	0.7288	0.7283	0.7647	1	1
12	BAT	T	2	198	99	99	0	0	0	0.5000	0.9268	0.9639	0.9705	0.9836	1	1
13	BLM	A	2	198	99	99	0	0	0	0.5000	0.9422	0.9688	0.9698	0.9903	1	1
14	Borrelia	T	2	198	99	99	0	0	0	0.5000	0.6022	0.6745	0.7338	0.7315	1	0
15	BPD	A	2	198	99	99	0	0	0	0.5000	0.9389	0.9739	0.9804	0.9928	1	1
16	BR	A	2	170	99	71	0	0	0	0.5824	0.7894	0.8860	0.9041	0.9367	1	1
17	Brucella abortus	T	2	180	99	81	0	0	0	0.5500	0.8075	0.8462	0.8231	0.8697	1	1
18	BSA	A	2	198	99	99	0	0	0	0.5000	0.8930	0.9705	0.9787	0.9971	1	1
19	BSE	A	2	198	99	99	0	0	0	0.5000	0.9465	0.9821	0.9865	0.9970	1	1
20	Ca	A	4	396	99	99	99	99	0	0.2500	0.5396	0.5952	0.6158	0.6619	1	1
21	CAD	A	2	198	99	99	0	0	0	0.5000	0.9122	0.9504	0.9575	0.9750	1	1

22	Callus	T	2	150	99	51	0	0	0	0.6600	0.7839	0.8732	0.8874	0.9167	1	1
23	CAM	A	2	198	99	99	0	0	0	0.5000	0.8142	0.9372	0.9326	0.9556	1	1
24	Cardiac pacemaker	T	2	198	99	99	0	0	0	0.5000	0.8327	0.8776	0.9065	0.9141	1	1
25	CCD	A	2	141	99	42	0	0	0	0.7021	0.9124	0.9770	0.9818	0.9979	1	1
26	CCl4	A	2	198	99	99	0	0	0	0.5000	0.8857	0.9593	0.9675	0.9877	1	1
27	CDA	A	2	198	99	99	0	0	0	0.5000	0.9298	0.9736	0.9796	0.9970	1	1
28	CDR	A	2	147	99	48	0	0	0	0.6735	0.8605	0.9521	0.9619	0.9824	1	1
29	Cell	AT	2	198	99	99	0	0	0	0.5000	0.8340	0.8809	0.9208	0.9323	1	1
30	Cement	T	2	185	99	86	0	0	0	0.5351	0.7523	0.7838	0.7951	0.8414	1	1
31	CH	A	2	148	91	57	0	0	0	0.6149	0.7719	0.8460	0.8592	0.8925	1	1
32	Cholera	T	2	198	99	99	0	0	0	0.5000	0.8161	0.8319	0.8612	0.8771	1	1
33	CI	A	2	183	99	84	0	0	0	0.5410	0.8175	0.8824	0.9199	0.9307	1	1
34	Cilia	T	2	156	99	57	0	0	0	0.6346	0.8446	0.8314	0.9100	0.9396	1	1
35	CIS	A	2	153	99	54	0	0	0	0.6471	0.8905	0.9621	0.9652	0.9871	1	1
36	CNS	A	2	198	99	99	0	0	0	0.5000	0.9197	0.9480	0.9572	0.9679	1	1
37	Coffee	T	2	198	99	99	0	0	0	0.5000	0.7110	0.6634	0.7593	0.7550	1	0
38	Cold	AT	3	260	99	99	62	0	0	0.3808	0.6538	0.7357	0.7800	0.8334	1	1
39	Compliance	T	2	198	99	99	0	0	0	0.5000	0.7227	0.7701	0.8370	0.8251	1	0
40	Cortex	T	2	198	99	99	0	0	0	0.5000	0.8728	0.9392	0.9515	0.9697	1	1
41	Cortical	T	3	297	99	99	99	0	0	0.3333	0.6587	0.7197	0.7830	0.8578	1	1
42	CP	A	3	297	99	99	99	0	0	0.3333	0.8494	0.9381	0.9468	0.9918	1	1
43	Crack	T	2	163	99	64	0	0	0	0.6074	0.8740	0.8970	0.9368	0.9396	1	1
44	CRF	A	2	198	99	99	0	0	0	0.5000	0.9414	0.9696	0.9774	0.9957	1	1
45	cRNA	A	2	198	99	99	0	0	0	0.5000	0.8803	0.8691	0.9576	0.9668	1	1
46	Crown	T	2	198	99	99	0	0	0	0.5000	0.6755	0.7545	0.8236	0.8119	1	0
47	CTX	A	2	183	99	84	0	0	0	0.5410	0.9497	0.9751	0.9769	0.9947	1	1
48	DAT	A	2	198	99	99	0	0	0	0.5000	0.8914	0.9550	0.9717	0.9931	1	1
49	DBA	A	2	183	99	84	0	0	0	0.5410	0.9319	0.9685	0.9721	0.9888	1	1
50	dC	A	2	198	99	99	0	0	0	0.5000	0.8508	0.9418	0.9454	0.9641	1	1
51	DDD	A	2	198	99	99	0	0	0	0.5000	0.8476	0.8839	0.8963	0.9129	1	1
52	DDS	A	3	220	99	99	22	0	0	0.4500	0.8252	0.8563	0.8935	0.9479	1	1
53	DE	A	2	126	99	27	0	0	0	0.7857	0.7911	0.8440	0.8143	0.8666	1	1
54	DI	A	2	198	99	99	0	0	0	0.5000	0.9459	0.9719	0.9740	0.9946	1	1
55	Digestive	T	2	198	99	99	0	0	0	0.5000	0.6733	0.7108	0.7696	0.7802	1	1
56	DON	A	2	126	99	27	0	0	0	0.7857	0.8682	0.9263	0.9383	0.9620	1	1
57	drinking	T	2	198	99	99	0	0	0	0.5000	0.7855	0.8773	0.8872	0.9155	1	1
58	eCG	A	2	198	99	99	0	0	0	0.5000	0.8990	0.9203	0.9507	0.9648	1	1
59	Eels	AT	2	130	99	31	0	0	0	0.7615	0.8834	0.9354	0.9276	0.9600	1	1
60	EGG	T	2	198	99	99	0	0	0	0.5000	0.6933	0.7275	0.7827	0.7768	1	0
61	EM	A	2	129	99	30	0	0	0	0.7674	0.8544	0.9639	0.9592	0.9877	1	1
62	EMS	A	2	198	99	99	0	0	0	0.5000	0.8615	0.8666	0.9372	0.9370	1	0
63	Epi	A	2	198	99	99	0	0	0	0.5000	0.8569	0.9214	0.9366	0.9725	1	1
64	ERP	A	2	198	99	99	0	0	0	0.5000	0.9388	0.9827	0.9867	0.9972	1	1
65	ERUPTION	T	2	197	99	98	0	0	0	0.5025	0.8952	0.9340	0.9504	0.9657	1	1
66	Erythrocytes	T	2	198	99	99	0	0	0	0.5000	0.6470	0.7023	0.7277	0.7292	1	1
67	Exercises	T	2	198	99	99	0	0	0	0.5000	0.6745	0.7064	0.7613	0.7742	1	1
68	FA	A	2	198	99	99	0	0	0	0.5000	0.8963	0.9704	0.9759	0.9969	1	1
69	Familial Adenomatous Polyposis	T	2	198	99	99	0	0	0	0.5000	0.7379	0.7624	0.8062	0.8045	1	0
70	FAS	A	2	198	99	99	0	0	0	0.5000	0.9421	0.9824	0.9850	0.9999	1	1
71	Fe	A	2	198	99	99	0	0	0	0.5000	0.8091	0.8625	0.8899	0.9089	1	1
72	Fish	AT	2	198	99	99	0	0	0	0.5000	0.8261	0.8943	0.9178	0.9316	1	1

73	Follicle	T	2	198	99	99	0	0	0	0.5000	0.8198	0.9057	0.9168	0.9531	1	1
74	Follicles	T	2	198	99	99	0	0	0	0.5000	0.8134	0.9236	0.9402	0.9634	1	1
75	FTC	A	2	198	99	99	0	0	0	0.5000	0.8557	0.9117	0.9482	0.9541	1	1
76	GAG	A	2	198	99	99	0	0	0	0.5000	0.9036	0.9507	0.9541	0.9815	1	1
77	Gamma-Interferon	T	2	198	99	99	0	0	0	0.5000	0.6857	0.7410	0.7775	0.7696	1	0
78	Ganglion	T	2	198	99	99	0	0	0	0.5000	0.8603	0.8679	0.8987	0.9088	1	1
79	Gas	T	2	198	99	99	0	0	0	0.5000	0.8199	0.8299	0.8892	0.9185	1	1
80	Glycoside	T	2	198	99	99	0	0	0	0.5000	0.8020	0.8929	0.8913	0.9409	1	1
81	Haemophilus ducreyi	T	2	153	99	54	0	0	0	0.6471	0.7798	0.8822	0.8803	0.8974	1	1
82	HCl	A	2	198	99	99	0	0	0	0.5000	0.9458	0.9779	0.9800	0.9943	1	1
83	Heregulin	T	2	173	99	74	0	0	0	0.5723	0.6726	0.7382	0.7880	0.7935	1	1
84	HGF	A	2	192	99	93	0	0	0	0.5156	0.7301	0.8180	0.8620	0.8727	1	1
85	HHV 8	A	2	176	99	77	0	0	0	0.5625	0.7612	0.8006	0.8015	0.8266	1	1
86	Hip	T	2	165	99	66	0	0	0	0.6000	0.7261	0.7694	0.7738	0.7981	1	1
87	HIV	A	2	198	99	99	0	0	0	0.5000	0.6927	0.7115	0.7801	0.7785	1	0
88	HPS	A	2	178	99	79	0	0	0	0.5562	0.9556	0.9808	0.9869	0.9987	1	1
89	HR	A	2	109	99	10	0	0	0	0.9083	0.9234	0.9407	0.9237	0.9455	1	1
90	Hybridization	T	2	198	99	99	0	0	0	0.5000	0.7879	0.8605	0.8744	0.8893	1	1
91	IA	A	2	134	99	35	0	0	0	0.7388	0.8322	0.9394	0.9388	0.9725	1	1
92	Ice	AT	3	235	99	99	37	0	0	0.4213	0.8193	0.8552	0.8882	0.9192	1	1
93	INDO	A	2	122	99	23	0	0	0	0.8115	0.8479	0.9536	0.9519	0.9646	1	1
94	Ion	T	2	198	99	99	0	0	0	0.5000	0.7443	0.7903	0.8290	0.8488	1	1
95	IP	A	2	196	99	97	0	0	0	0.5051	0.9259	0.9715	0.9793	0.9941	1	1
96	Iris	T	2	161	99	62	0	0	0	0.6149	0.8202	0.8748	0.8946	0.9036	1	1
97	ITP	A	2	198	99	99	0	0	0	0.5000	0.8694	0.9562	0.9679	0.9868	1	1
98	JP	A	2	192	99	93	0	0	0	0.5156	0.9216	0.9644	0.9541	0.9834	1	1
99	LABOR	T	2	198	99	99	0	0	0	0.5000	0.7472	0.7902	0.8427	0.8495	1	1
100	Lactation	T	2	198	99	99	0	0	0	0.5000	0.7977	0.8406	0.8690	0.8873	1	1
101	Language	T	2	198	99	99	0	0	0	0.5000	0.7540	0.8186	0.8870	0.9014	1	1
102	Laryngeal	T	2	198	99	99	0	0	0	0.5000	0.6726	0.7485	0.7869	0.7828	1	0
103	Lawsonia	T	2	115	99	16	0	0	0	0.8609	0.8670	0.9379	0.9327	0.9606	1	1
104	Leishmaniasis	T	2	161	99	62	0	0	0	0.6149	0.8008	0.8388	0.8723	0.8908	1	1
105	lens	T	3	297	99	99	99	0	0	0.3333	0.7036	0.7406	0.8007	0.7972	1	0
106	Lupus	T	3	297	99	99	99	0	0	0.3333	0.6730	0.6730	0.7804	0.7269	1	0
107	lymphogranulomatosis	T	2	119	99	20	0	0	0	0.8319	0.8508	0.8832	0.8782	0.9073	1	1
108	MAF	A	2	120	99	21	0	0	0	0.8250	0.8855	0.9641	0.9618	0.9809	1	1
109	Malaria	T	2	198	99	99	0	0	0	0.5000	0.7610	0.7971	0.8565	0.8457	1	0
110	MBP	A	2	143	99	44	0	0	0	0.6923	0.7505	0.9089	0.9132	0.9444	1	1
111	MCC	A	2	131	99	32	0	0	0	0.7557	0.8668	0.9883	0.9863	0.9987	1	1
112	Medullary	T	2	198	99	99	0	0	0	0.5000	0.7700	0.8250	0.8689	0.9029	1	1
113	MHC	A	2	198	99	99	0	0	0	0.5000	0.8971	0.9563	0.9630	0.9849	1	1
114	Milk	T	2	198	99	99	0	0	0	0.5000	0.7486	0.8117	0.8192	0.8394	1	1
115	Moles	T	2	174	99	75	0	0	0	0.5690	0.7859	0.8247	0.8841	0.8863	1	1
116	MRS	A	2	166	99	67	0	0	0	0.5964	0.9511	0.9781	0.9792	0.9947	1	1
117	Murine sarcoma virus	T	2	180	99	81	0	0	0	0.5500	0.6753	0.7140	0.7384	0.7330	1	0
118	NBS	A	2	146	99	47	0	0	0	0.6781	0.9000	0.9783	0.9786	0.9962	1	1
119	NEUROFIBROMATOSIS	T	2	198	99	99	0	0	0	0.5000	0.7170	0.7519	0.7912	0.8068	1	1
120	NM	A	2	122	84	38	0	0	0	0.6885	0.8112	0.8596	0.9208	0.9233	1	1
121	NPC	A	2	163	99	64	0	0	0	0.6074	0.9627	0.9877	0.9897	0.9999	1	1

122	Nurse	T	2	198	99	99	0	0	0	0.5000	0.6490	0.7057	0.7733	0.7770	1	1
123	Nursing	T	2	198	99	99	0	0	0	0.5000	0.7276	0.7085	0.8124	0.7620	1	0
124	OCD	A	2	198	99	99	0	0	0	0.5000	0.8744	0.9683	0.9707	0.9963	1	1
125	OH	A	2	198	99	99	0	0	0	0.5000	0.8200	0.9095	0.9378	0.9586	1	1
126	Orf	AT	2	198	99	99	0	0	0	0.5000	0.8706	0.8529	0.9343	0.9427	1	1
127	ORI	A	2	123	99	24	0	0	0	0.8049	0.8677	0.9377	0.9580	0.9858	1	1
128	PAF	A	2	115	99	16	0	0	0	0.8609	0.9021	0.9853	0.9887	0.9935	1	1
129	Parotitis	T	2	198	99	99	0	0	0	0.5000	0.6952	0.7630	0.8079	0.8393	1	1
130	PCA	A	5	491	99	99	99	99	95	0.2016	0.7591	0.8475	0.8942	0.9685	1	1
131	PCB	A	2	127	99	28	0	0	0	0.7795	0.8675	0.9570	0.9585	0.9797	1	1
132	PCD	A	2	198	99	99	0	0	0	0.5000	0.9266	0.9758	0.9786	0.9946	1	1
133	PCP	A	3	297	99	99	99	0	0	0.3333	0.8599	0.9081	0.9388	0.9781	1	1
134	PEP	A	2	198	99	99	0	0	0	0.5000	0.8578	0.9492	0.9616	0.9787	1	1
135	PHA	A	2	110	99	11	0	0	0	0.9000	0.9077	0.9537	0.9423	0.9671	1	1
136	Pharmaceutical	T	2	198	99	99	0	0	0	0.5000	0.7823	0.8408	0.8791	0.8874	1	1
137	Phosphorus	T	2	198	99	99	0	0	0	0.5000	0.6658	0.7387	0.7942	0.8032	1	1
138	Phosphorylase	T	2	166	99	67	0	0	0	0.5964	0.7338	0.8082	0.8094	0.8110	1	1
139	pI	A	2	156	99	57	0	0	0	0.6346	0.8934	0.9620	0.9744	0.9862	1	1
140	Plague	T	2	168	99	69	0	0	0	0.5893	0.7600	0.8260	0.8421	0.8568	1	1
141	Plaque	T	2	197	99	98	0	0	0	0.5025	0.8845	0.9480	0.9646	0.9799	1	1
142	Platelet	T	2	198	99	99	0	0	0	0.5000	0.6823	0.7262	0.7813	0.8058	1	1
143	Pleuropneumonia	T	2	198	99	99	0	0	0	0.5000	0.8113	0.8626	0.8821	0.9014	1	1
144	Pneumocystis	T	2	198	99	99	0	0	0	0.5000	0.7095	0.8060	0.8233	0.8141	1	0
145	POL	A	2	162	99	63	0	0	0	0.6111	0.8479	0.9433	0.9585	0.9680	1	1
146	Polymyalgia Rheumatica	T	2	198	99	99	0	0	0	0.5000	0.7737	0.8482	0.8603	0.8874	1	1
147	posterior pituitary	T	2	194	99	95	0	0	0	0.5103	0.7746	0.7819	0.8278	0.8462	1	1
148	Potassium	T	2	198	99	99	0	0	0	0.5000	0.7268	0.7600	0.8043	0.8140	1	1
149	PR	A	2	165	99	66	0	0	0	0.6000	0.8230	0.9445	0.9546	0.9753	1	1
150	Projection	T	2	198	99	99	0	0	0	0.5000	0.8067	0.8747	0.8816	0.9249	1	1
151	PVC	A	2	198	99	99	0	0	0	0.5000	0.8983	0.9591	0.9656	0.9886	1	1
152	RA	A	3	297	99	99	99	0	0	0.3333	0.8622	0.9066	0.9310	0.9753	1	1
153	Radiation	T	2	198	99	99	0	0	0	0.5000	0.6912	0.7148	0.8056	0.7878	1	0
154	RB	A	2	198	99	99	0	0	0	0.5000	0.8814	0.9388	0.9550	0.9709	1	1
155	RBC	A	2	198	99	99	0	0	0	0.5000	0.6938	0.7217	0.7788	0.7892	1	1
156	rDNA	A	2	198	99	99	0	0	0	0.5000	0.7227	0.7524	0.8241	0.8297	1	1
157	Respiration	T	2	198	99	99	0	0	0	0.5000	0.7297	0.7980	0.8576	0.8672	1	1
158	Retinal	T	2	198	99	99	0	0	0	0.5000	0.7211	0.7494	0.8222	0.8274	1	1
159	Root	T	2	198	99	99	0	0	0	0.5000	0.8688	0.8758	0.8966	0.9277	1	1
160	RSV	A	2	134	99	35	0	0	0	0.7388	0.8528	0.9446	0.9470	0.9703	1	1
161	SARS-associated coronavirus	T	2	118	71	47	0	0	0	0.6017	0.8301	0.8727	0.8906	0.8865	1	0
162	SARS	A	2	198	99	99	0	0	0	0.5000	0.8622	0.8838	0.9280	0.9314	1	1
163	SCD	A	2	198	99	99	0	0	0	0.5000	0.8977	0.9576	0.9677	0.9946	1	1
164	Schistosoma mansoni	T	2	198	99	99	0	0	0	0.5000	0.7351	0.7491	0.8037	0.7978	1	0
165	Semen	T	2	186	99	87	0	0	0	0.5323	0.7584	0.8414	0.8649	0.8816	1	1
166	sex factor	T	2	131	96	35	0	0	0	0.7328	0.7872	0.8621	0.9003	0.9115	1	1
167	SLS	A	2	164	99	65	0	0	0	0.6037	0.9353	0.9880	0.9865	1.0000	1	1
168	Sodium	T	2	197	99	98	0	0	0	0.5025	0.7279	0.7597	0.7756	0.7810	1	1
169	SPR	A	2	198	99	99	0	0	0	0.5000	0.9490	0.9777	0.9821	0.9984	1	1

170	SS	A	2	144	98	46	0	0	0	0.6806	0.9184	0.9808	0.9779	0.9990	1	1
171	Staph	T	2	198	99	99	0	0	0	0.5000	0.7318	0.7142	0.7804	0.7906	1	1
172	STEM	AT	2	198	99	99	0	0	0	0.5000	0.9048	0.9337	0.9448	0.9657	1	1
173	Sterilization	T	2	198	99	99	0	0	0	0.5000	0.7234	0.7703	0.8188	0.8441	1	1
174	Strep	T	2	197	99	98	0	0	0	0.5025	0.7526	0.7887	0.8114	0.8080	1	0
175	Synapsis	T	2	134	99	35	0	0	0	0.7388	0.8501	0.8972	0.9022	0.9142	1	1
176	TAT	A	3	297	99	99	99	0	0	0.3333	0.6961	0.7548	0.7840	0.7848	1	1
177	Tax	AT	2	180	99	81	0	0	0	0.5500	0.8749	0.8856	0.9295	0.9317	1	1
178	TEM	A	2	198	99	99	0	0	0	0.5000	0.8499	0.8857	0.9357	0.9777	1	1
179	THYMUS	T	3	297	99	99	99	0	0	0.3333	0.7493	0.7525	0.8136	0.8252	1	1
180	TLC	A	2	198	99	99	0	0	0	0.5000	0.9157	0.9650	0.9764	0.9885	1	1
181	TMJ	A	2	198	99	99	0	0	0	0.5000	0.6671	0.6738	0.7507	0.7359	1	0
182	TMP	A	2	150	99	51	0	0	0	0.6600	0.7815	0.8864	0.9057	0.9308	1	1
183	TNC	A	2	167	99	68	0	0	0	0.5928	0.9324	0.9730	0.9752	0.9966	1	1
184	TNT	A	2	198	99	99	0	0	0	0.5000	0.9233	0.9718	0.9677	0.9939	1	1
185	Tolerance	T	2	198	99	99	0	0	0	0.5000	0.7822	0.8322	0.8491	0.8758	1	1
186	tomography	T	2	198	99	99	0	0	0	0.5000	0.7762	0.7738	0.8205	0.8380	1	1
187	Torula	T	2	122	88	34	0	0	0	0.7213	0.7997	0.8394	0.8457	0.8487	1	1
188	TPA	A	2	198	99	99	0	0	0	0.5000	0.8937	0.9372	0.9581	0.9773	1	1
189	TPO	A	2	198	99	99	0	0	0	0.5000	0.8636	0.9302	0.9487	0.9738	1	1
190	TRF	A	2	179	99	80	0	0	0	0.5531	0.9105	0.9404	0.9552	0.9730	1	1
191	TYR	A	2	198	99	99	0	0	0	0.5000	0.7728	0.8776	0.8937	0.8991	1	1
192	US	A	2	198	99	99	0	0	0	0.5000	0.7600	0.8024	0.8895	0.9152	1	1
193	Ventricles	T	2	198	99	99	0	0	0	0.5000	0.7659	0.8668	0.8975	0.9145	1	1
194	veterinary	T	2	198	99	99	0	0	0	0.5000	0.6474	0.6785	0.7061	0.6790	1	0
195	Wasp	AT	2	198	99	99	0	0	0	0.5000	0.9095	0.9200	0.9504	0.9691	1	1
196	WBS	A	2	128	93	35	0	0	0	0.7266	0.8542	0.9586	0.9407	0.9872	1	1
197	WT1	A	2	198	99	99	0	0	0	0.5000	0.7220	0.7063	0.7708	0.7730	1	1
198	Yellow Fever	T	2	183	99	84	0	0	0	0.5410	0.7270	0.8293	0.8684	0.8773	1	1

Table A2. Interactive learning results for 74 ambiguous abbreviations in the UMN corpus. Please see the caption of Table A1 for explanation of the header.

ID	Word	Type	#S	#inst	#inst in top 5 senses					major sense ratio	ALC scores				IL > AL	IL > RR
					S1	S2	S3	S4	S5		RS	AL	RR	IL		
1	AB	A	11	499	345	137	8	2	1	0.6914	0.7117	0.7296	0.7310	0.8478	1	1
2	VBG	A	2	500	299	201	0	0	0	0.5980	0.8961	0.9449	0.9446	0.9622	1	1
3	AC	A	11	500	161	158	118	42	9	0.3220	0.6691	0.6959	0.7053	0.8040	1	1
4	ALD	A	5	500	407	88	3	1	1	0.8140	0.8731	0.9212	0.9160	0.9542	1	1
5	AMA	A	3	500	444	31	25	0	0	0.8880	0.8798	0.9137	0.9259	0.9499	1	1
6	ASA	A	3	500	404	93	3	0	0	0.8080	0.9080	0.9374	0.9378	0.9560	1	1
7	AVR	A	7	500	381	103	5	4	4	0.7620	0.7635	0.8105	0.7975	0.9073	1	1
8	AV	A	4	500	374	116	8	2	0	0.7480	0.7678	0.8054	0.8054	0.8663	1	1
9	BAL	A	2	500	457	43	0	0	0	0.9140	0.8973	0.9392	0.9446	0.9704	1	1
10	BK	A	2	500	343	157	0	0	0	0.6860	0.8222	0.9244	0.9422	0.9550	1	1
11	BMP	A	4	500	456	36	7	1	0	0.9120	0.8443	0.8484	0.8406	0.9026	1	1
12	BM	A	4	500	459	25	14	2	0	0.9180	0.8855	0.9005	0.8914	0.9381	1	1
13	C&S	A	5	500	434	47	16	2	1	0.8680	0.9410	0.9783	0.9784	0.9832	1	1
14	C3	A	4	500	249	243	6	2	0	0.4980	0.8442	0.8622	0.8887	0.9262	1	1
15	C4	A	5	500	261	231	6	1	1	0.5220	0.8058	0.8479	0.8540	0.9046	1	1
16	CA	A	4	500	391	105	2	2	0	0.7820	0.7840	0.8139	0.8332	0.8570	1	1
17	CDI	A	4	500	270	225	3	2	0	0.5400	0.8780	0.9224	0.9168	0.9655	1	1
18	CEA	A	5	500	444	53	1	1	1	0.8880	0.8423	0.8732	0.8772	0.9240	1	1
19	CR	A	6	500	453	28	16	1	1	0.9060	0.9122	0.9315	0.9315	0.9443	1	1
20	CTA	A	5	500	396	100	2	1	1	0.7920	0.8832	0.9249	0.9246	0.9661	1	1
21	CVA	A	2	500	278	222	0	0	0	0.5560	0.9133	0.9516	0.9470	0.9653	1	1
22	CVP	A	3	500	436	62	2	0	0	0.8720	0.8705	0.9296	0.9141	0.9603	1	1
23	CVS	A	3	500	457	41	2	0	0	0.9140	0.9255	0.9699	0.9595	0.9784	1	1
24	DC	A	8	500	282	152	31	31	1	0.5640	0.6331	0.6714	0.6981	0.7834	1	1
25	DIP	A	3	500	462	36	2	0	0	0.9240	0.9262	0.9590	0.9613	0.9827	1	1
26	DM	A	5	500	286	209	3	1	1	0.5720	0.8098	0.8552	0.8656	0.9128	1	1
27	DT	A	8	500	336	129	23	4	3	0.6720	0.6923	0.7470	0.7365	0.8373	1	1
28	EC	A	5	499	439	45	11	2	2	0.8798	0.8945	0.9137	0.9065	0.9370	1	1
29	ER	A	3	500	448	34	18	0	0	0.8960	0.8996	0.9240	0.9345	0.9539	1	1
30	ES	A	6	500	469	14	8	7	1	0.9380	0.8394	0.8359	0.8315	0.9355	1	1
31	ET	A	8	500	289	200	6	1	1	0.5780	0.7731	0.8389	0.8289	0.9287	1	1
32	FISH	AT	2	500	449	51	0	0	0	0.8980	0.9152	0.9772	0.9748	0.9889	1	1
33	FSH	A	3	500	265	231	4	0	0	0.5300	0.7603	0.8267	0.8345	0.8691	1	1
34	GT	A	6	500	446	30	16	5	2	0.8920	0.8479	0.8648	0.8527	0.9094	1	1
35	IA	A	9	500	275	176	19	11	5	0.5500	0.7305	0.7704	0.7564	0.8725	1	1
36	IB	A	9	500	472	8	8	5	2	0.9440	0.8470	0.8604	0.8619	0.9477	1	1
37	IM	A	3	500	461	38	1	0	0	0.9220	0.8934	0.9286	0.9269	0.9604	1	1
38	IR	A	5	500	394	102	2	1	1	0.7880	0.8695	0.9047	0.9043	0.9522	1	1
39	IT	AT	12	500	225	103	58	48	40	0.4500	0.5951	0.6022	0.6329	0.7505	1	1
40	IVF	A	4	500	308	188	3	1	0	0.6160	0.8594	0.8785	0.8922	0.9182	1	1
41	LA	A	6	500	426	40	30	2	1	0.8520	0.8719	0.9085	0.9055	0.9441	1	1
42	LE	A	9	500	345	134	5	5	3	0.6900	0.7220	0.8070	0.7870	0.8951	1	1
43	MOM	A	4	500	439	57	3	1	0	0.8780	0.9683	0.9864	0.9798	0.9908	1	1
44	MP	A	14	500	179	107	105	55	12	0.3580	0.3826	0.4144	0.4161	0.5390	1	1
45	MR	A	6	500	314	176	5	3	1	0.6280	0.7695	0.8079	0.8141	0.8990	1	1
46	MSSA	A	2	500	418	82	0	0	0	0.8360	0.8539	0.9239	0.9142	0.9426	1	1

47	MS	A	10	500	279	207	4	3	2	0.5580	0.6551	0.6979	0.7002	0.7818	1	1
48	NAD	A	2	500	377	123	0	0	0	0.7540	0.8469	0.8805	0.9071	0.9102	1	1
49	NA	A	5	500	474	14	10	1	1	0.9480	0.9644	0.9759	0.9726	0.9856	1	1
50	NP	A	6	500	438	53	5	2	1	0.8760	0.8601	0.9044	0.8987	0.9508	1	1
51	OP	A	8	500	308	121	55	6	5	0.6160	0.8660	0.8938	0.8815	0.9519	1	1
52	OR	AT	4	500	466	32	1	1	0	0.9320	0.9160	0.9412	0.9427	0.9665	1	1
53	OTC	A	2	500	469	31	0	0	0	0.9380	0.9292	0.9502	0.9259	0.9534	1	1
54	PAC	A	10	500	275	137	47	25	7	0.5500	0.6971	0.7157	0.7340	0.8519	1	1
55	PA	A	8	500	212	138	83	61	2	0.4240	0.6909	0.7359	0.7415	0.8531	1	1
56	PCP	A	5	500	294	111	93	1	1	0.5880	0.6967	0.7432	0.7638	0.7977	1	1
57	PDA	A	3	500	361	138	1	0	0	0.7220	0.7618	0.8402	0.8705	0.8849	1	1
58	PD	A	15	500	409	34	14	9	8	0.8180	0.6612	0.6786	0.6822	0.8848	1	1
59	PE	A	4	500	408	89	2	1	0	0.8160	0.7921	0.8656	0.8715	0.9256	1	1
60	PM	A	4	500	423	74	2	1	0	0.8460	0.8290	0.9001	0.8718	0.9363	1	1
61	PR	A	7	500	252	141	88	12	4	0.5040	0.8883	0.9081	0.9073	0.9566	1	1
62	PT	A	5	500	455	22	21	1	1	0.9100	0.8822	0.9077	0.9116	0.9473	1	1
63	RA	A	5	500	394	66	36	3	1	0.7880	0.7666	0.8142	0.8103	0.8618	1	1
64	RT	A	8	500	336	149	7	2	2	0.6720	0.8177	0.8475	0.8402	0.9190	1	1
65	SA	A	7	498	373	88	29	4	2	0.7490	0.8506	0.8968	0.9078	0.9357	1	1
66	SBP	A	2	500	417	83	0	0	0	0.8340	0.8495	0.8920	0.9098	0.9102	1	1
67	SMA	A	6	500	353	84	56	3	2	0.7060	0.6944	0.7316	0.7389	0.8262	1	1
68	SS	A	3	500	439	57	4	0	0	0.8780	0.9767	0.9843	0.9803	0.9857	1	1
69	T1	A	6	500	198	194	103	3	1	0.3960	0.6762	0.6843	0.7335	0.7642	1	1
70	T2	A	7	500	227	166	97	7	1	0.4540	0.6620	0.7060	0.7184	0.7721	1	1
71	T3	A	6	500	268	156	65	5	4	0.5360	0.7294	0.7980	0.7970	0.8710	1	1
72	T4	A	3	500	424	41	35	0	0	0.8480	0.8407	0.8959	0.9079	0.9374	1	1
73	US	AT	4	500	402	94	3	1	0	0.8040	0.8439	0.8952	0.9035	0.9131	1	1
74	VAD	A	5	500	396	87	13	3	1	0.7920	0.7662	0.7912	0.8145	0.8765	1	1

Table A3. Interactive learning results for 24 ambiguous abbreviations in the VUH corpus. Please see the caption of Table A1 for explanation of the header.

ID	Word	Type	#S	#inst	#inst in top 5 senses					major sense ratio	ALC scores				IL > AL	IL > RR
					S1	S2	S3	S4	S5		RS	AL	RR	IL		
1	ad	AT	9	200	181	6	4	3	2	0.9050	0.7820	0.7535	0.7877	0.8490	1	1
2	ag	A	3	171	117	51	3	0	0	0.6842	0.7668	0.7938	0.8198	0.8040	1	0
3	bm	A	7	199	128	54	11	2	2	0.6432	0.6803	0.6704	0.7207	0.7551	1	1
4	ca	A	6	200	128	37	19	8	7	0.6400	0.8377	0.8544	0.8728	0.8451	0	0
5	cc	A	6	200	114	32	30	18	4	0.5700	0.8150	0.7913	0.8374	0.8792	1	1
6	cm	A	2	200	199	1	0	0	0	0.9950	0.9748	0.9724	0.9760	0.9815	1	1
7	dm	A	2	200	170	30	0	0	0	0.8500	0.7908	0.8174	0.8526	0.8745	1	1
8	gtt	A	4	200	143	46	9	2	0	0.7150	0.8460	0.8429	0.8787	0.8643	1	0
9	hd	A	8	199	112	79	3	1	1	0.5628	0.4951	0.4892	0.4911	0.5209	1	1
10	hs	A	6	191	147	20	13	8	2	0.7696	0.7337	0.7369	0.7619	0.8128	1	1
11	icd	A	2	199	195	4	0	0	0	0.9799	0.9732	0.9661	0.9795	0.9388	0	0
12	lad	A	4	200	150	48	1	1	0	0.7500	0.8824	0.9013	0.9142	0.9376	1	1
13	le	A	3	200	178	14	8	0	0	0.8900	0.9093	0.9190	0.9112	0.9365	1	1
14	ln	A	3	144	136	4	4	0	0	0.9444	0.8999	0.9092	0.9100	0.9570	1	1
15	med	A	5	195	96	79	12	6	2	0.4923	0.6887	0.6898	0.7093	0.7364	1	1
16	mg	A	2	200	197	3	0	0	0	0.9850	0.9734	0.9753	0.9761	0.9834	1	1
17	mi	A	2	200	199	1	0	0	0	0.9950	0.9563	0.9492	0.9565	0.9894	1	1
18	pe	A	5	200	65	61	53	16	5	0.3250	0.6781	0.6853	0.7251	0.7503	1	1
19	pt	A	5	198	179	9	6	2	2	0.9040	0.7895	0.7519	0.8036	0.8462	1	1
20	ra	A	4	200	149	36	14	1	0	0.7450	0.8623	0.8762	0.8856	0.9029	1	1
21	si	A	3	185	168	16	1	0	0	0.9081	0.8890	0.9080	0.9107	0.9308	1	1
22	sle	A	3	185	178	6	1	0	0	0.9622	0.9244	0.8794	0.9351	0.8895	1	0
23	ss	A	6	196	116	47	27	3	2	0.5918	0.8511	0.8820	0.9056	0.9411	1	1
24	tia	A	2	200	199	1	0	0	0	0.9950	0.9457	0.9268	0.9366	0.9551	1	1