Classifying Medical Questions based on an Evidence Taxonomy

Hong Yu[1], Carl Sable[2], Hai Ran Zhu[3]

[1]Columbia University
Department of Biomedical Informatics
622 West, 168th Street, VC-5, NY, NY 10032
Hong.Yu@dbmi.columbia.edu

[2]Cooper Union
Department of Electrical and Computer Engineering
51 Astor Place, NY, NY 10003
sable2@cooper.edu

[3]Columbia University
Department of Computer Science
500 West, 120th Street, NY, NY 10021
Hz2116@cs.columbia.edu

Abstract

We present supervised machine-learning approaches to automatically classify medical questions based on a hierarchical evidence taxonomy created by physicians. We show that SVMs is the best classifier for this task and that a ladder approach, which incorporates the knowledge representation of the hierarchical evidence taxonomy, leads to the highest performance. We have explored the use of features from a large, robust biomedical knowledge resource, namely, the Unified Medical Language System (UMLS), and we have found that performance is generally enhanced by including these features in addition to bag-of-words.

Introduction

Question classification is a task that assigns a given question to one or more predefined question categories. In the context of open-domain question answering, question taxonomies are created so that specific answer strategies can be developed for specific question types defined in the taxonomy (Harabagiu et al. 2000) (Hovy et al. 2001). Typical approaches for question classification combine surface cues (e.g., "What" and "When") with named entities (e.g., PERSON and LOCATION) (Hovy et al. 2001). In the medical domain, on the other hand, we found that physicians have classified questions based on medical, domain-specific knowledge (Bergus et al. 2000) (Ely et al. 2000, 2002). The purpose of classifying medical questions goes beyond the development of question-type answer generation strategies.

In the medical domain, physicians are urged to practice *Evidence Based Medicine* when faced with questions about how to care for their patients (Gorman et al., 1994) (Straus and Sackett, 1999) (Bergus et al., 2000). Evidence based medicine refers to the use of the best evidence from scientific and medical research to make decisions about the care of individual patients. Physicians are trained to ask "good" questions, such as "What is the drug of choice for epididymitis?", that require answers which incorporate clinical studies. On the other hand, the question "What is causing her anemia" is considered a "bad" question because it is patient-specific and therefore unanswerable (Ely et al. 2002). Ely and his colleagues have created an *Evidence Taxonomy* (in Figure 1) to categorize questions that were potentially answerable with evidence (Ely et al. 2002). These researchers have concluded that only *Evidence* questions are potential answerable with evidence using medical literature and other online medical resources.

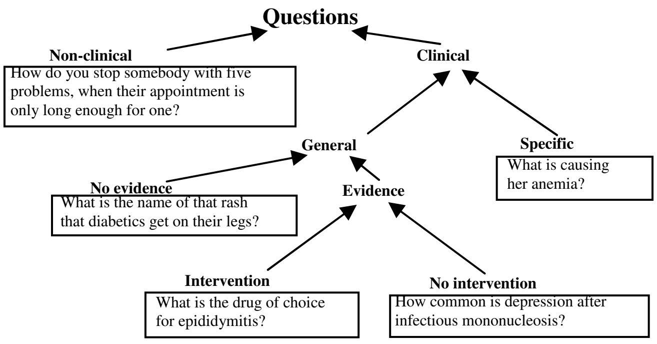
Questions
Non-clinical
How do you stop somebody with five problems, when their appointment is only long enough for one?
Clinical
General
Specific
What is causing her anemia?
No evidence
What is the name of that rash that diabetics get on their legs?
Evidence
Intervention
What is the drug of choice for epididymitis?
No intervention
How common is depression after infectious mononucleosis?

Figure 1: "Evidence taxonomy" created by Ely and his colleagues (Ely et al. 2002) with examples.

Previously, we have developed approaches to automatically separate answerable questions from unanswerable ones (Yu and Sable 2005). This study focuses on the harder task of automatically classifying questions to the specific categories presented in the evidence taxonomy; namely, *Clinical* vs *Non-Clinical*, *General* vs *Specific*, *Evidence* vs *No Evidence*, and *Intervention* vs *No Intervention*. We utilize a large biomedical knowledge resource, namely, the Unified Medical Language System (UMLS), to aid our automatic question classification.

Question classification based on an evidence-taxonomy has practical importance. Such a classification may be useful for advising medical students, residents, and physicians to formulate "good questions"; i.e., questions that can be answered with evidence. For example, when a physician in training asks the patient-specific question, "What is causing her anemia?", we may provide the physician advice to reformulate the question as "What causes anemia spherocytic?", "What causes anemia macrocytic?", or "What causes anemia folic acid deficiency?", from which the physician who posed the question might choose one for further answer generation.

In addition, similar to the role of question classification in the context of open-domain question answering, evidence-based question classification is useful for generating question-type-specific answer strategies. For example, *No Intervention* questions generally belong to the family of factoid questions, for which short answers are usually expected. On the other hand, *Intervention* questions are generally scenario-based and require long and complex answers.

This study is a part of an effort to build a medical, domain-specific question answering system (Yu et al. 2005). In the following, we first describe the UMLS. We then describe, in detail, the question collection we have used for training and testing, as well as various supervised machine-learning systems that we have used for automatic classification. We then report our results, and end with a final discussion and conclusions.

The Unified Medical Language System

The Unified Medical Language System (UMLS)[1] (Humphreys and Lindberg 1993) links concepts that come from more than 100 terminologies, classifications, and thesauri, some in multiple editions, to aid in the development of computer systems that process text in the biomedical domain. The UMLS incorporates the Metathesaurus, a large database that currently incorporates more than one million biomedical concepts, plus synonyms and concept relations. For example, the UMLS links the following synonymous terms as a single concept: *Achondroplasia*, *Chondrodystrophia, Chondrodystrophia fetalis*, and *Osteosclerosis congenita.*

The UMLS additionally consists of the Semantic Network, which is a hierarchical semantic representation that contains 135 nodes or semantic types (e.g., *Pharmacologic Substance*); each semantic type represents a category to which certain UMLS concepts can be mapped. Each UMLS concept in the Metathesaurus is assigned one or more semantic types. For example, *Arthritis* is assigned to one semantic type, *Disease or Syndrome*; *Achondroplasia* is assigned to two semantic types, *Disease or Syndrome* and *Congenital Abnormality*.

Furthermore, the National Library of Medicine has made available MMTx[2], a programming implementation of MetaMap (Aronson 2001), which maps free text to UMLS concepts and their associated semantic types. The MMTx program first parses text, separating the text into noun phrases. Each noun phrase is then mapped to a set of possible UMLS concepts, taking into account spelling and morphological variations, and each concept is weighted, with the highest weight representing the most likely mapped concept. The UMLS concepts are then mapped to semantic types according to definitive rules as described in the previous paragraph. MMTx can be used either as a standalone application or as an API that allows systems to incorporate its functionality. In our study, we have applied and empirically evaluated MMTx to map terms in a question to appropriate UMLS concepts and semantic types; we have experimented with adding the resulting concepts and semantic types as additional features for question classification. Related work that uses UMLS concepts to improve document retrieval in the medical domain can be found in (Mao and Chu 2002).

Question Corpus and Categories

Ely and his colleagues (Ely et al. 2002) have annotated 200 questions that were randomly selected from over one thousand questions posed by family physicians (Ely et al. 1999) to be *Non-clincial* (8), *Specific* (57), *No-evidence* (18), *Intervention* (85), or *No-intervention* (32). We have used these 200 annotated medical questions for training and test sets. Note that *Intervention* is the largest category in this question collection, with 85 questions. Therefore, a baseline system that classifies every question into the largest category (i.e., *Intervention*) would achieve an overall accuracy of 42.5% and random guessing would lead to an overall accuracy of 20.0%.

[1] http://www.nlm.nih.gov/research/umls/

[2] http://mmtx.nlm.nih.gov/

Supervised Machine Learning

We have considered the classification of medical questions to be a standard text categorization task. We have explored various supervised machine-learning approaches to automatically assign labels to questions based on the evidence taxonomy. We present our results for the binary classifications *Clinical* vs *Non-Clinical*, *General* vs *Specific*, *Evidence* vs *No Evidence*, and *Intervention* vs *No Intervention*, as well as our results for five-way classification into the leaf categories *Non-clinical, Specific, No Evidence, Intervention,* and *No Intervention.* In the following subsections, we will describe the machine-learning systems, the learning features including bag of words and features selected from the Unified Medical Language System (UMLS), our training and testing methodology, and the evaluation metrics used for our classification.

Systems

We have applied seven supervised machine-learning approaches, namely, Rocchio/TF*IDF, K-nearest neighbors (kNN), maximum entropy, probabilistic indexing, naive Bayes, support vector machines, and BINS. All of these systems have been used successfully for text categorization tasks (Sebastiani 2002) (Sable 2003). For SVMs, we have used the implementation of libsvm[3]. For the rest of machine-learning systems except for BINs (Sable and Church 2001), we have used the implementation available in Rainbow[4] (McCallum 1996).

Rocchio/TF*IDF (Rocchio 1971) adopts TF*IDF, the vector space model typically used for information retrieval, for text categorization tasks. Rocchio/TF*IDF represents every document and category as a vector of TF*IDF values. The term frequency (TF) of a token (typically a word) is the number of times that the token appears in the document or category, and the inverse document frequency (IDF) of a token is a measure of the token's rarity (usually calculated based on the training set). For test documents, scores are assigned to each potential category by computing the similarity between the document to be labeled and the category, often computed using a simple cosine metric; the category with the highest score is then chosen.

K-Nearest Neighbors (kNN) determines which training questions are the most similar to each test question, and then uses the known labels of these similar training questions to predict a label for the test question. The similarity between two questions can be computed as the number of overlapping features between them, as the inverse of the Euclidean Distance between feature vectors, or according to other measures. The kNN approach has been successfully applied to a variety of text categorization tasks (Sebastiani, 2002) (Yang and Liu 1999).

Naïve Bayes is commonly used in machine learning and text categorization. Naïve Bayes is based on Bayes' Law and assumes conditional independence of features. For text categorization, this “naive” assumption amount to the assumption that the probability of seeing one word in a document is independent of the probability of seeing any other word in a document, given a specific category. Although this is clearly not true in reality, Naive Bayes has been useful for many text classification and other information retrieval tasks (Lewis 1998). The label of a question is the category that has the highest probability given the “bag of words” in the document. To be computationally feasible, log likelihood is generally maximized instead of probability.

Probabilistic Indexing is another probabilistic approach that chooses the category with the maximum probability given the words in a document. Probabilistic indexing stems from Fuhr’s probabilistic indexing paradigm (Fuhr 1988), which was originally intended for relevance feedback and was generalized for text categorization by Joachims (Joachims 1997), who considered it a probabilistic version of a TF*IDF classifier, although it more closely resembles Naive Bayes. Unlike Naive Bayes, the number of times that a word occurs in a document comes into play, because the probability of choosing each specific word, if a word were to be randomly selected from the document in question, is used in the probabilistic calculation. Although this approach is less common in the text categorization literature, one author of this paper has seen that it is very competitive for many text categorization tasks (Sable 2003). In this study, we show that probabilistic indexing generally outperforms all other machine-learning systems except for BINS and SVMs, and for certain tasks we consider, it performs the best of all systems.

Maximum Entropy is another probabilistic approach that has been successfully applied to text categorization (Nigam, Lafferty, and McCallum 1999). A maximum entropy system starts with the initial assumption that all categories are equally likely. It then iterates through a process known as improved iterative scaling that updates the estimated probabilities until some stopping criterion is met. After the process is complete, the category with the highest probability is selected.

Support Vector Machines (SVMs) is a binary classifier that learns a hyperplane in a feature space that acts as an optimal linear separator which separates (or nearly separates) a set of positive examples from a set of negative examples with the maximum possible margin (the margin is defined as the distance from the hyperplane to the closest of the positive and negative examples). SVMs have been

[3] Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] Available at http://www-2.cs.cmu.edu/~mccallum/bow/

widely tested to be one of the best machine-learning classifiers, and previous studies have shown that SVMs outperform other machine learning algorithms for open-domain sentence classification (Zhang and Lee 2003) and other text categorization tasks (Yang and Liu 1999) (Sebastiani 2002). Our study has shown that SVMs generally has the best performance for classifying medical questions based on the evidence taxonomy.

BINS (Sable and Church 2001) is a generalization of Naive Bayes. BINS places words that share common features into a single bin. Estimated probabilities of a token appearing in a document of a specific category are then calculated for bins instead of individual words, and this acts as a method of smoothing which can be especially important for small data sets. BINS has proven to be very competitive for many text categorization tasks (Sable 2003) (Yu and Sable 2005).

Classification and Ladder Approach

We report the results of classifying questions into binary classes based on the evidence taxonomy; namely, *Clinical* vs *Non-clinical, General* vs *Specific, Evidence* vs *No evidence,* and *Intervention* vs *No intervention,* by applying different machine learning systems.

In addition, we also apply the machine-learning systems to classify the questions into one of the five leaf-node categories of the evidence taxonomy, namely, *Non-clinical, Specific, No evidence, Intervention,* and *No intervention.* For this task, in addition to standard multi-class categorization by different machine-learning systems, we have experimented with a "ladder" approach that can only be used for hierarchical classification. This technique has proven useful for certain text categorization tasks involving a hierarchical taxonomy in the past (Koller and Sahami 1997) (Chakrabart et. al. 1998) (Ruiz and Srinivasan 1999) (Dumais and Chen 2000).

The ladder approach, sometimes referred to as a cascade of classifiers, utilizes the knowledge representation of the evidence taxonomy. For this task, the ladder approach performs five-class categorization by combining four independent binary classifications. It first predicts whether a question is *Clinical* vs *Non-clinical*. If a question is *Clinical*, it then predicts the question to be *General* vs *Specific.* If *General,* it further predicts to be *Evidence* vs *No evidence.* Finally, if *Evidence*, it classifies the question to be either *Intervention* or *No intervention.*

Learning Features

Since our collection consists of medical, domain-specific questions, we have incorporated biomedical terminology from the robust, state of the art Unified Medical Language System (UMLS) as additional learning features for question classification. In our study, we apply bag of words as a baseline, and add in (or substitute) UMLS concepts and semantic types as additional features. Since we apply MMTx, the tool that maps question strings to appropriate UMLS concepts and semantic types, we first evaluate the performance of MMTx and then report the results of our question classification.

Cross-Validation and Evaluation Metrics

To evaluate the performance of each system, we have performed four-fold cross-validation. We have randomly divided our data set into four sets of categories consisting of 50 questions each. We have used one set for testing and the other set for training, and we have repeated such an experiment testing each set of questions once. We apply four-fold cross-validation instead of the typical ten-fold because the number of training set in one category (i.e., *Non-clinical*) is eight, which is less than ten. In addition, we have experimented with ten-fold cross validations and found that the results are comparable to four-fold. We therefore report four-fold only. We report overall accuracy, which is simply the percentage of questions that are categorized correctly.

Results and Evaluation

Since we have applied MMTx for identifying appropriate UMLS concepts and semantic types for each question, which are then included as features for question classification, we have evaluated the precision of MMTx for this task. One of the authors (Dr. Carl Sable) has manually examined the 200 questions comprising our corpus as described in earlier. MMTx assigns 769 UMLS Concepts and 924 semantic types to the 200 questions (recall that some UMLS concepts are mapped to more than one semantic type, as described in the UMLS section). Our analysis has indicated that 164 of the UMLS Concept labels and 194 of the semantic type labels are wrong; this indicates precisions of 78.7% and 79.0%, respectively. An example of a case that MMTx gets wrong is the abbreviation "pt", which, in this corpus, is often used as an abbreviation for "patient"; MMTx typically assigns this to the UMLS concept *pint* and the semantic type *Quantitative Concept*. Note that a manual estimation of the recall of MMTx would be difficult, since it would require an expert that is familiar with all possible UMLS concepts and the various ways to express them.

Table 1 shows the results in terms of overall accuracy for all seven systems applied to each of our binary categories using various combinations of features. The baseline systems that classify every question into the largest category would achieve an overall accuracy of 96.0%, 70.3%, 86.7%, and 72.6% for *Clinical* vs *Non-clinical, General* vs *Specific, Evidence* vs *No Evidence,* and *Intervention* vs *No Intervention.* All results that beat these baselines are in bold, and results that tie these baselines are underlined.

Table 1: Binary classification results (overall accuracy %) using different classifiers and features (**bold** indicates the accuracy is above the baseline and underscore indicates the accuracy is equal to the baseline; "Average" reports the average accuracy of all seven systems; "Best" reports the highest accuracy of the seven systems)

ML Approach	**Performance Using Various Features (C means UMLS Concepts, ST means semantic types)**					
	Bag of Words	**Words+C**	**Words+ST**	**Words+C+ST**	**C only**	**ST only**
Rocchio/TF*IDF						
Clinical* vs *Non-clinical	90.5	86.5	94.0	91.0	93.0	82.5
General* vs *Specific	**74.5**	68.8	**74.5**	**72.9**	69.3	63.0
Evidence* vs *No Evidence	76.3	85.2	81.5	85.2	83.0	82.2
Intervention* vs *No Intervention	70.1	69.2	69.2	72.6	66.7	63.2
K-Nearest Neighbors						
Clinical* vs *Non-clinical	95.0	95.5	96.0	96.0	96.0	96.0
General* vs *Specific	**77.1**	**71.2**	**75.5**	**76.0**	**74.0**	**71.9**
Evidence* vs *No Evidence	83.7	85.2	86.7	86.7	85.9	86.7
Intervention* vs *No Intervention	72.6	**76.1**	71.8	72.6	**74.4**	71.8
Naive Bayes						
Clinical* vs *Non-clinical	96.0	95.5	96.0	95.0	96.0	96.0
General* vs *Specific	**79.2**	70.3	**76.0**	**74.0**	**74.5**	**71.9**
Evidence* vs *No Evidence	85.2	83.7	85.2	83.7	86.7	85.9
Intervention* vs *No Intervention	72.6	**77.8**	72.6	**76.1**	**74.4**	66.7
Probabilistic Indexing						
Clinical* vs *Non-clinical	91.0	87.0	94.0	91.0	86.0	83.0
General* vs *Specific	**78.1**	**73.4**	**75.5**	**78.1**	**74.0**	69.3
Evidence* vs *No Evidence	78.5	79.3	82.2	83.7	81.5	82.2
Intervention* vs *No Intervention	**74.4**	**73.5**	**77.8**	**78.6**	**74.4**	**77.8**
Maximum Entropy						
Clinical* vs *Non-clinical	94.0	93.0	94.0	95.0	94.5	93.0
General* vs *Specific	**75.5**	70.3	**76.0**	**74.0**	**71.4**	68.6
Evidence* vs *No Evidence	80.7	83.0	83.7	85.9	82.2	80.7
Intervention* vs *No Intervention	70.1	70.9	73.5	71.8	72.6	65.0
Support Vector Machines						
Clinical* vs *Non-clinical	96.0	96.0	96.0	96.0	96.0	96.0
General* vs *Specific	**76.6**	**77.1**	**79.7**	**79.7**	**76.6**	**75.0**
Evidence* vs *No Evidence	**88.1**	**87.4**	**88.9**	**88.9**	**87.4**	**88.9**
Intervention* vs *No Intervention	**74.4**	**76.9**	**75.2**	**76.1**	**75.2**	**73.5**
BINS						
Clinical* vs *Non-clinical	92.5	92.0	91.0	92.5	94.0	93.5
General* vs *Specific	**78.1**	**78.6**	**77.6**	**74.5**	**71.9**	64.6
Evidence* vs *No Evidence	**87.4**	**88.9**	85.9	**88.1**	**87.4**	84.4
Intervention* vs *No Intervention	70.9	**76.1**	**74.4**	**76.9**	71.8	72.6
Average						
Clinical* vs *Non-clinical	93.6	92.2	94.4	93.8	93.6	91.4
General* vs *Specific	**77.0**	**72.8**	**76.4**	**75.6**	**73.1**	69.2
Evidence* vs *No Evidence	82.8	84.7	84.9	86.0	84.9	84.4
Intervention* vs *No Intervention	72.2	**74.4**	**73.5**	**75.0**	**72.8**	70.1
Best						
Clinical* vs *Non-clinical	96.0	96.0	96.0	96.0	96.0	96.0
General* vs *Specific	**79.2**	**78.6**	**79.7**	**79.7**	**76.6**	**75.0**
Evidence* vs *No Evidence	**88.1**	**88.9**	**88.9**	**88.9**	**87.4**	**88.9**
Intervention* vs *No Intervention	**74.4**	**77.8**	**77.8**	**78.6**	**75.2**	**77.8**

Based on the performance scores in Table 1, it seems clear that SVMs outperforms all other systems in most of cases. Every result ties or beats the baseline. For three of the four classification tasks, SVMs is responsible for the highest classification accuracies; that is, 96.0% for *Clinical* vs *Non-clinical*, 79.7% for *General* vs *Specific*, and 88.9% for *Evidence* vs *No Evidence* using a combination of all features or words plus semantic types. For the *Intervention* vs *No Intervention* task, the best result is achieved by the probabilistic indexing system, which achieves a 78.6% overall accuracy using a combination of all features. In general, the two most promising systems after SVMs are the probabilistic indexing system and BINS. Interestingly, our earlier results show that probabilistic indexing has out-performed other machine-learning systems (e.g., SVMs and BINS) for classifying a question as either *Answerable* or *Unanswerable* (Yu and Sable 2005). The probabilistic indexing system performs better than SVMs for the *Intervention* vs *No Intervention* task, and BINS is the only system apart from SVMs that beats the baseline for half of the task/feature combinations (the probabilistic indexing system beats the baseline for just under half of the task/feature combinations). The naïve Bayes and kNN systems also both beat or tie the baseline for the majority of cases.

We have found that including the UMLS concepts and semantic types as additional features in general enhances the performance of all classifiers. For example, when classifying between *General* and *Specific,* SVMs performs at 79.7% when including the UMLS features; the accuracy is 3.1% higher than with the bag-of-words only. On the other hand, the degree to which UMLS features contribute to the performance could be subtle. For example, in case of *Evidence* vs *No Evidence,* SVMs enhances only 0.8% to the overall of 88.9% accuracy when including the UMLS features in addition to the bag-of-words. The results of applying the UMLS features only (i.e., without bag-of-words), are mixed. For example, when using semantic types as the only features for SVMs learning, the classifier outperforms the bag-of-words for *Evidence* vs *No Evidence,* but underperforms the bag-of-words for *General* vs *Specific.*

We have created the bottom two sections of Table 1 to summarize the results of all classifiers. The first of these two sections shows the average accuracy of all seven systems for each binary classification task with each feature combination. The second of these two sections shows the best accuracy of the seven systems for each classification task with each feature combination. Note that 17 of the 24 best results match the corresponding result of the SVMs system (at times tied with other systems). Of the other seven best results, three come from the probabilistic indexing system, two come from the BINS system, and two come from the Naïve Bayes system. Furthermore, the addition of these two bottom sections of the table, showing the average and best overall accuracies, respectively, makes it simpler to compare the effects of various features. Based on the results of average accuracy, we have found that using either UMLS concepts or semantic types *instead of* bag of words sometimes helps and sometimes hurts. Typically, the results using either of these features are about the same as using bag of words, the standard feature to use in text categorization; these results support the claim that they are useful. For the *General* versus *Specific* classification task, however, performance degrades by several percentage points. Using these features *in addition to* bag of words improves performance in five out of eight cases. Two of the three cases in which it does not help involve the *General* versus *Specific* task. Finally, adding both new features to bag of words leads to improvement for three out of four tasks (the exception again being *General* versus *Specific*), and for the *Evidence* versus *No Evidence* task and the *Intervention* versus *No Intervention* task, the improvement is several percentage points. Looking at the best performances, we see that the column reporting results for experiments combining all features contains the best results for every one of the classification tasks (at times tied with results from other columns).

Recall that for five-way classification, we have experimented with two approaches. The first approach is a standard, multi-class *flat* categorization, in which each classifier is trained with the training sets consisting of documents with five labels; in this case, *Non-clinical, Specific, No evidence, Intervention,* and *No intervention.* The ladder approach utilizes the knowledge representation of the evidence taxonomy. The predicted label for a document depends on hierarchical inference from four independent, binary classifiers that are trained for *Non clinical* vs *Clinical, General* vs *Specific, Evidence* vs *No evidence,* and *Intervention* vs *No intervention.* The results of the ladder approach are shown in the top section of Table 2, with the results for flat classification at the bottom of Table 2. All results in either section that are better than the corresponding result in the other section are in bold; results that are equal in the two sections are underlined.

Table 2 shows the results in terms of overall accuracy for all seven systems applied to the entire set of the five leaf categories from the hierarchy using all combinations of features. Recall that a baseline system that classifies every question into the largest category (i.e., *Intervention*) would achieve an overall accuracy of 42.5%, and random guessing would achieve an overall accuracy of 20.0%. Our results show that all of our systems, using either of the two approaches previously explained, beat the baseline with almost all of the available feature combinations (the exceptions are all in the last two columns, representing experiments that replaced bag of words with UMLS features). The highest overall accuracy is 59.5%, which is 17% above the baseline and represents a 30.1% error

Table 2: Five category classification results (overall accuracy %) using a standard, flat category classification approach (bottom) and the ladder approach (top) (**bold** indicates a result better than the other approach, underscore indicates equal results)

Ladder	**Performance Using Various Features (C means UMLS Concepts, ST means semantic types)**					
	Bag of Words	**Words+C**	**Words+ST**	**Words+C+ST**	**C only**	**ST only**
Rocchio/TF*IDF	**49.0**	45.0	48.5	50.5	41.5	34.5
kNN	50.5	49.0	48.0	49.0	**50.5**	45.0
Naïve Bayes	**52.0**	**50.5**	51.0	52.5	**51.0**	**46.5**
Prob Indexing	**53.0**	50.5	53.5	**57.5**	47.0	45.0
MaxEnt	**48.0**	46.0	**52.5**	**50.5**	**48.5**	43.0
SVMs	**54.0**	**57.0**	**59.5**	**58.0**	**55.5**	**57.0**
BINS	52.5	**56.0**	**53.0**	**53.0**	**49.0**	40.5
Standard	**Performance Using Various Features (C means UMLS Concepts, ST means semantic types)**					
	Bag of Words	**Words+C**	**Words+ST**	**Words+C+ST**	**C only**	**ST only**
Rocchio/TF*IDF	47.5	45.0	**50.5**	**52.0**	**45.5**	**41.0**
kNN	**52.5**	**51.0**	48.0	**50.5**	49.5	**46.5**
Naïve Bayes	51.5	50.0	**51.5**	**53.5**	50.0	45.0
Prob Indexing	52.0	**51.0**	**56.5**	57.0	**47.5**	**45.5**
MaxEnt	46.0	**47.5**	48.0	49.5	47.0	**45.0**
SVMs	51.0	53.5	54.0	53.0	52.0	50.5
BINS	**53.0**	51.0	51.5	48.0	47.5	40.5

reduction; this is achieved by the SVMs system using a combination of bag of words and semantic types and a ladder approach. The best accuracy from a flat categorization approach is 57.0%; this is achieved by the probabilistic indexing system using a combination of all features.

Our results show that SVMs' ladder approach beats the standard five-way flat classification by a 3~7% enhancement in accuracy with all six feature conditions. Furthermore, SVMs ladder approach outperforms all other classifiers with either a standard, flat or a ladder approach, and with different features. Our results strongly support that the knowledge of the category taxonomy is useful for question classification and that using SVMs with a ladder approach leads to the best classifier. These results do not come with a surprise because SVMs is intrinsically a binary classifier, and the binary-status of the steps of the ladder makes it more appropriate for SVM classification.

On the other hand, the ladder approach applied to the other classifiers has shown mixed results; the performance of the ladder approach beats the standard approach for 22 task/feature combinations (including all six of the results for SVMs), ties the standard approach for 3 task/feature combinations, and underperforms the standard approach for 17 task/feature combinations. Although it is not entirely clear when the ladder approach is better than the standard approach, it is important that the ladder approach has helped SVMs, which has clearly performed the best overall, and this has led to our highest overall accuracy for these categories.

Looking back at the results from Table 1, we might notice that the systems do not perform equally well on all binary classification tasks. Recall that SVMs performs the best for three of the four tasks, but for one task (*Intervention* vs *No Intervention*), probabilistic indexing performs the best. It is therefore possible that the ladder approach might achieve better results if separate classifiers were used at different steps of the ladder.

Conclusions and Future Work

This paper presents our results of automatic classification of medical questions using various supervised machine-learning systems. For this study, we have used a small data set consisting of 200 questions that have been manually labeled by experts into five mutually exclusive, hierarchical categories. We have experimented with both binary classification corresponding to the individual decisions in the hierarchy, and also classification into the five leaf categories. Classification into the leaf categories has been performed using a standard approach and also the ladder approach. The performance measures for our binary classification tasks, in terms of overall accuracy, were all reasonably high. The performance was not generally high for classification into five categories, although almost all systems with all sets of features did beat the baseline. We speculate that the performance would improve if we were to obtain more training data; for example, a previous study that applies SVMs to open-domain sentence classification has shown that the performance of SVMs for a five-class categorization task increases from 68.0% to 80.2% when the total number of questions for training increases from 1000 to 5500 (Zhang and Lee, 2003).

We found that including the UMLS concepts and semantic types as additional features can enhance results, although it does not do so in all cases. Even using these novel features in place of bag of words tends to achieve performance close to bag of words. Certainly these features look promising for future work involving larger training sets.

Our primary future task is to automatically classify questions involving medical knowledge, and to use the classification in order to decompose biomedical questions according to component question types, as proposed by (Ely et al. 2000) and (Bergus et al, 2000). Each component question type represents a specific category for a question indicating what type of information is sought; for example, "What are the affects of <*drug*> on <*disease*>?". If questions can be accurately classified into component question types, one significant advantage of this would be that specific answer strategies could be developed for each question type. We speculate that the recognition of UMLS concepts and semantic types, along with deeper natural language processing and analysis of questions, will also play a crucial role in this type of question classification.

Acknowledgements: We thank Dr. John Ely for valuable discussions and for making the annotated medical questions available to us. This project is supported by the Juvenile Diabetes Research Foundation International (JDRF 6-2005-835). Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect JDRF's views.

References

Aronson, A. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proceedings of the American Medical Information Association Symposium.

Bergus, GR., Randall, CS, Sinift, SD, and Rosenthal, DM. 2000. Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? *Arc Fam Med* 9 (541-547).

Chakrabarti, S.; B. Dom; R. Agrawal; and P. Raghavan. 1998. Scalable feature selection, classification, and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB Journal* 7(3):163-178.

Dumais, S. and H. Chen. 2000. Hierarchical classification of web content. In Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

Ely, J.; Osheroff, J.; Ebell, M.; Bergus, G.; Barcey, L.; Chambliss, M.; and Evans, E. 1999. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal* 319:358-361.

Ely, J.; Osheroff, J.; Gorman, P.; Ebell, M.; Chambliss, M.; Pifer, E.; and Stavri, P. 2000. A taxonomy of generic clinical questions: classification study. *British Medical Journal* 321:429-432.

Ely, J.; Osheroff, J.; Ebell, M.; Chambliss, M.; Vinson, D.; Stevermer, J.; and Pifer, E. 2002. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *British Medical Journal* 324:710-713.

Fuhr, N. 1998. Models for retrieval with probabilistic indexing. *Information Processing and Management* 25(1):55-72.

Gorman, P.; Ash, J.; and Wykoff, L. 1994. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association* 82(2):140-146.

Harabagiu, S, Pasca, M, and Maiorano S. 2000. FALCON: Boosting knowledge for answer engines. In *Proceeding of the 9th text retrieval conference (TREC-9).* NIST.

Hovy, EH, Gerber, L, Hermjakob, U, Lin, CY, and Ravichandran, D. 2001. Toward semantics-based answer pinpointing. In Proceedings of the Human Language Technology Conference (HLT2001).

Humphreys, B. and Lindberg, D. 1993. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association* 81: 170-7.

Jacquemart, P. and Zweigenbaum, P. 2003. Towards a medical question-answering system: a feasibility study. *Studies in Health Technology and Informatics* 95:463-8.

Joachims, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of the 14th International Conference on Machine Learning.

Koller, D. and M. Sahami. 1997. Hierarchically classifying documents using very few words. In Proceedings of the 14th International Conference on Machine Learning.

Lewis, D. 1998. Naive (Bayes) at forty: the independence assumption in information retrieval. In Proceedings of the European Conference on Machine Learning.

Mao W. and W. Chu. 2002. Free-text medical document retrieval via phrase-based vector space model. Proceedings of the American Medical Information Association Symposium.

McCallum, A. 1996. *A toolkit for statistical language modeling, text retrieval, classification, and clustering. http://www.cs.cmu.edu/~mccallum/bow.*

Mosteller, F. and D. Wallace. 1963. Inference in an authorship problem. Journal of the American Statistical Association 58:275-309.

Nigam, K.; Lafferty, J.; and McCallum, A. 1999. Using maximum entropy for text classification. In Proceedings of the IJCAI-99 workshop on machine learning for information filtering.

Rocchio, J. 1971. Relevance feedback in information retrieval. In The Smart Retrieval System: Experiments in

Automatic Document Processing, pages 313-323, Prentice Hall.

Ruiz, M. E. and P. Srinivasan. 1999. Hierarchical neural networks for text categorization. In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval.

Sable, C. 2003. Robust Statistical Techniques for the Categorization of Images Using Associated Text. Ph.D. dissertation, Columbia University.

Sable, C. and K. Church. 2001. Using Bins to empirically estimate term weights for text categorization. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing.

Sackett, D.; Straus, S.; Richardson, W.; Rosenberg, W.; and Haynes, R. 2000. *Evidence-Based Medicine: How to practice and teach EBM*. Edinburgh: Harcourt Publishers Limited.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34:1-47.

Suzuki, J.; Taira, H.; Sasaki Y.; and Maeda, E. 2003. Question classification using HDAG kernel. In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. pp. 61-68.

Straus, S. and Sackett, D. 1999. Bringing evidence to the point of care. *Journal of the American Medical Association* 281:1171-1172.

van Rigsbergen, C. J. 1979. *Information Retrieval, 2nd Edition*. London: Butterworths.

Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In Proceedings in the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Yu, H. and Sable, C. 2005. *Being Erlang Shen: Identifying Answerable Questions*. To appear in *Nineteenth International Joint Conference on Artificial Intelligence Workshop on Knowledge and Reasoning for Answering Question.*.

Yu, Y.; Zhu, H.; Sable, C.; Shapiro, S.; Osheroff, J.; Ely, J.; and Cimino, J. 2005. Beyond Information Retrieval - Biomedical Question Answering. Forthcoming.

Zhang, D. and Lee, W. S. 2003. Question classification using support vector machines. In Proceedings of the 26th Annual International ACM SIGIR conference, pages 26-32.