# Shallow Information Extraction from Medical Forum Data

**Parikshit Sondhi** **Manish Gupta** **ChengXiang Zhai** **Julia Hockenmaier**
University of Illinois at Urbana Champaign
{sondhi1, gupta58, czhai, juliahmr}@illinois.edu

## Abstract

We study a novel shallow information extraction problem that involves extracting sentences of a given set of topic categories from medical forum data. Given a corpus of medical forum documents, our goal is to extract two related types of sentences that describe a biomedical case (i.e., medical problem descriptions and medical treatment descriptions). Such an extraction task directly generates medical case descriptions that can be useful in many applications. We solve the problem using two popular machine learning methods Support Vector Machines (SVM) and Conditional Random Fields (CRF). We propose novel features to improve the accuracy of extraction. Experiment results show that we can obtain an accuracy of up to 75%.

## 1 Introduction

Conventional information extraction tasks generally aim at extracting finer granularity semantic information units such as entities and relations. While such detailed information is no doubt very useful, extraction of such information also tends to be difficult especially when the mentions of the entities to be extracted do not conform to regular syntactic patterns.

In this paper, we relax this conventional goal of extraction and study an easier extraction task where we aim at extracting sentences that belong to a set of predefined semantic categories. That is, we take a sentence as a unit for extraction. Specifically, we study this problem in the context of extracting medical case description from medical forums.

A variety of medical health forums exist online. People use them to post their problems, get advices from experienced patients, get second opinions from other doctors, or merely to vent out their frustration.

Compared with well-structured sources such as Wikipedia, forums are more valuable in the sense that they contain first hand patient experiences with richer information in terms of what treatments are better than others and why. Besides this, on forums, patients explain their symptoms much more freely than those mentioned on relatively formal sources like Wikipedia. And hence, forums are much more easier to understand for a naïve user.

However, even on targeted forums (which focus on a single disease), data is quite unstructured. There is therefore a need to structure out this information and present it in a form that can directly be used for a variety of other information extraction applications like the collecting of medical case studies pertaining to a particular disease, mining frequently discussed symptoms, identifying correlation between symptoms and treatments, etc.

A typical medical case description tends to consist of two aspects:

- **Physical Examination/Symptoms (PE)**: This covers current conditions and includes any condition that is the focus of current discussion. Note that if a drug causes an allergy, then we consider it as a PE and not a medication. Any condition that is the focus of conversation, i.e. around which treatments are being proposed or questions are

being asked is considered PE even if the user is recounting their past experience.

- **Medications (MED)**: Includes medications the person is currently taking, or is intending to take, or any medication on which the question is targeted. Medications do not necessarily mean drugs. Any measures (including avoiding of substances) taken to treat or avoid the symptoms are considered as medication. Sometimes, users also mention other things like constituents of the drug, how much of the drug to consume at a time, how to get access to a medication, how much it costs, side effects of medications, other qualities of medications etc.

Figure 1 shows an example of PE and MED labelings.

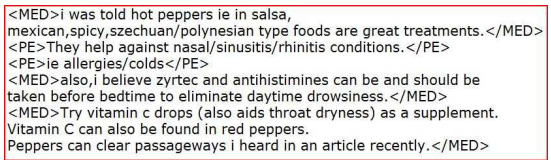


Figure 1: Example of PE and MED labelings

We thus frame the problem of extracting medical case descriptions as extracting sentences that describe any of these two aspects. Specifically, the task is to identify sentences in each of the two related categories (i.e., PE and MED) from forum posts. As an extraction task, this task is "shallower" than conventional information extraction tasks such as entity extraction in the sense that we extract a sentence as a unit, which makes the extraction task more tractable. Indeed, the task is more similar to sentence categorization. However, it also differs from a regular sentence categorization task (e.g., sentiment analysis) in that the multiple categories are usually closely related and categorization of multiple sentences may be dependent in the sense that knowing the category of one sentence may influence our decision about the category of another sentence nearby. For example, knowing that a sentence is in the category PE should increase our belief that the next sentence is of category of PE or MED.

We solve the problem using two popular machine learning methods, Support Vector Machines (SVM) and Conditional Random Fields (CRF). We define and study a large set of features, including two kinds of novel features: (1) novel features based on semantic generalization of terms, and (2) novel features specific to forums.

Since this is a novel task, there is no existing data set that we can use for evaluation. We thus create a new data set for evaluation. Experiment results show that both groups of novel features are effective and can improve extraction accuracy. With the best configurations, we can obtain an accuracy of up to 75%, demonstrating feasibility of automatic extraction of medical case descriptions from forums.

## 2 Related work

Medical data mining has been looked atleast since the early 2000s. Cios and Moore (2002) emphasize the uniqueness of medical data mining. They stress that data mining in medicine is distinct from that in other fields, because the data are heterogeneous, and special ethical, legal, and social constraints apply to private medical information. Treatment recommendation systems have been built that use the structured data to diagnose based on symptoms (Lazarus et al., 2001) and recommend treatments. Holt et al.(2005) provide references to medical systems that use case based reasoning methodologies for medical diagnosis. Huge amounts of medical data stored in clinical data warehouses can be used to detect patterns and relationships, which could provide new medical knowledge (Lazarus et al., 2001). In contrast, we look at the problem of converting some of the unstructured medical text data present in forum threads into structured symptoms and treatments. This data can then be used by all of the above mentioned applications.

Structuring of unstructured text has been studied by many works in the literature. Automatic information extraction (Aone and Ramos-Santacruz, 2000; Buttler et al., 2001) and wrapper induction techniques have been used for structuring web data. Sarawagi (2008) and Laender et al. (2002) offer comprehensive overviews of information extraction and wrapper induction

techniques respectively. The main difference between our work and main stream work on extraction is that we extract sentences as units, which is shallower but presumably more robust. Heinze et al. (2002) state that the current state-of-the-art in NLP is suitable for mining information of moderate content depth across a diverse collection of medical settings and specialties. Zhou et al. (2006), the authors perform information extraction from clinical medical records using a decision tree based classifier using resources such as WordNet [1], UMLS [2] etc. They extract past medical history and social behaviour from the records.

In other related works, sentiment classification (Pang et al., 2002; Prabowo and Thelwall, 2009; Cui et al., 2006; Dave et al., 2003) attempts to categorize text based on polarity of sentiments and is often applied at the sentence level (Kim and Zhai, 2009). Some work has also been done on extracting content from forum data. This includes finding question answer pairs (Cong et al., 2008) from online forums, auto-answering queries on a technical forum (Feng et al., 2006), ranking answers (Harabagiu and Hickl, 2006) etc. To the best of our knowledge, this is the first work on shallow extraction from medical forum data.

## 3 Problem formulation

Let $P = (s_1, ...s_n)$ be a sequence of sentences in a forum post. Given a set of interesting categories $C = \{c_1, ..., c_k\}$ that describe a medical case, our task is to extract sentences in each category from the post $P$. That is, we would like to classify each sentence $s_i$ into one of the categories $c_i$ or *Background*, which we treat as a special category meaning that the sentence is irrelevant to our extraction task. Depending on specific applications, a sentence may belong to more than one category.

In this paper, we focus on extracting sentences of two related categories describing a medical case: (1) Physical Examination (PE), which includes sentences describing the condition of a patient (i.e., roughly symptoms) (2) Medications (MED), which includes sentences mentioning medications (i.e., roughly treatment). These sentences provide a basic description of a medical case and can already be very useful if we can extract them.

We chose to analyze at the sentence level because a sentence provides enough context to detect the category accurately. For example, detecting the categories at word level will not help us to mark a sentence like *"I get very uncomfortable after eating cheese"* as PE or mark a sentence like *"It's best to avoid cheese in that case"* as MED. Here the problem is loosely represented by a combination of *"uncomfortable eating cheese"* and the solution is represented loosely by *"avoid cheese"*. Indeed, in preliminary analysis, we found that most of the times, the postings consist of PE and MED type sentences.

## 4 Methods

We use SVMs and CRFs to learn classifiers to solve our problem. SVMs represent approaches that solve the problem as a classification/categorization task while CRFs solve the problem as a sequence labeling task. In this section, we provide the basics of SVMs and CRFs.

### 4.1 Support Vector Machines

SVM first introduced in (Boser et al., 1992), are a binary classifier that constructs a hyperplane which separates the training instances belonging to the two classes. SVMs maximize the separation margin between this hyperplane and the nearest training datapoints of any class. The larger the margin, the lower the generalization error of the classifier. SVMs have been used to classify both linearly and non-linearly seperable data, and have been shown to outperform other popular classifiers like decision trees, Naïve Bayes classifiers, k-nearest neighbor classifiers, etc. We use SVMs as a representative classifier that does not consider dependencies between the predictions on multiple sentences.

### 4.2 Conditional Random Fields

Each of the sentences in the postings can itself contain features which help us to categorize it.

[1] http://wordnet.princeton.edu/

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

Besides this, statistical dependencies exist between sentences. Intuitively, a MED sentence will follow a PE sentence with high probability, but the probability of a PE sentence following an MED sentence would be low. Conditional random fields are graphical models that can capture such dependencies among input sentences. A CRF model defines a conditional distribution $p(y|x)$ where $y$ is the predicted category (label) and $x$ is the set of sentences (observations). CRF is an undirected graphical model in which each vertex represents a random variable whose distribution is to be inferred, and each edge represents a dependency between two random variables. The observation $x$ can be dependent on the current hidden label $y$, previous $n$ hidden labels and on any of the other observations in a $n$ order CRF. CRFs have been shown to outperform other probabilistic graphical models like Hidden Markov Models (HMMs) and Maximum Entropy Markov Models (MeMMs). Sutton and McCallum (2006) provide an excellent tutorial on CRFs.

## 5 Features

To perform our categorization task, we use the following features.

- **Word based features**: This includes unigrams, bigrams and trigrams in the current sentence. Each of the n-grams is mapped to a separate boolean feature per sentence where value is 1 if it appears in sentence and 0 otherwise.

- **Semantic features**: This includes Unified Medical Language System (UMLS[3]) semantic groups of words in the current sentence. UMLS is a prominent bio-medical domain ontology. It contains approximately a million bio-medical concepts grouped under 135 semantic groups. MMTX[4] is a tool that allows mapping of free text into UMLS concepts and groups. We use these 135 semantic groups as our semantic features. In order to generate these features, we first process this sentence through MMTX API which provides all the semantic groups that were found in the sentence. Each of the semantic groups becomes a boolean feature.

- **Position based features**: We define two types of position based features: position of the current sentence in the post and position of the current post in the thread. These features are specific to the forum data. We include these features based on the observations that first post usually contains condition related sentences while subsequent posts often contain treatment measures for the corresponding condition. Each of the position number of a sentence in a post and a post in a thread is mapped to a boolean feature which gets fired for a sentence at a particular position. E.g. For a sentence at position $i$ in a post, POSITION_IN_POST_$i$ would be set to 1 while other features POSITION_IN_POST_$j$ where $j \neq i$ would be set to 0.

- **User based features**: We include a boolean feature which gets fired when the sentence is a part of a post by the thread creator. This feature is important because most of the posts by a thread creator have a high probability of being a PE.

- **Tag based features(Edge features)**: We define features on tags (PE/MED/Backgnd) of previous two sentences to capture local dependencies between sentences. E.g., a set of medication related tags often follow a description of a condition. We use these features only for CRF based experiments.

- **Morphological features**: These include one boolean feature each for presence of
  - a capitalized word in the sentence
  - an abbreviation in the sentence
  - a number in the sentence
  - a question mark in the sentence
  - an exclamation mark in the sentence

- Length based features: We also consider the number of words in a sentence as a separate type of feature. Feature LENGTH_$i$

[3] http://www.nlm.nih.gov/research/umls/
[4] http://mmtx.nlm.nih.gov/

| Category | Labeler 1 | Labeler 2 |
|---|---|---|
| PE | 513 | 517 |
| MED | 286 | 280 |
| Background | 695 | 697 |

Table 1: Labeling results

becomes true for a sentence containing $i$ words.

## 6 Experiments

### 6.1 Dataset

Evaluation of this new extraction task is challenging as no test set is available. To solve this problem, we opted to created our own test set. HealthBoards[5] is a medical forum web portal that allows patients to discuss their ailments. We scraped 175 posts contained in 50 threads on allergy i.e., an average of 3.5 posts per thread and around 2 posts per user with a maximum of 9 posts by a particular user. Two humans were asked to tag this corpus as conditions (i.e., PE category) or treatments (i.e., MED category) or none on a per sentence basis. The corpus consists of 1494 sentences. Table 1 shows the labeling results. The data set is available at (http://timan.cs.uiuc.edu/downloads.html). Also the labeling results match quite well (82.86%) with a Kappa statistic value of 0.73. Occasionally (around 3%) PE and MED both occur in the same sentence and the labelers chose to mark such sentences as PE. In the case when the two labelers disagree, we manually analyzed the results and further chose one of them for our experiments.

### 6.2 Evaluation methodology

For evaluation, we use 5-fold cross validation. For CRFs, we used the Mallet[6] toolkit and for SVM, we used SVM-Light[7]. We experimented by varying the size of the training set, with different feature sets, using two machine learning models: SVMs and CRFs. Our aim is to accurately classify any sentence in a post as PE or MED or background. First we explore and identify the feature sets that help us in attaining higher accuracy. Next, we identify the setting (sequence labeling by CRFs or independent classification by SVMs) that works better to model our problem. We present most of our results using four metrics: precision, recall, F1 measure and average accuracy which is the ratio of correctly labeled sentences to the total sentences.

We considered the following features: all the 2647 words in the vocabulary (no stop-word removal or any other type of selection), 10858 bigrams, 135 semantic groups from UMLS, two position based features, one user based feature, two tag based features, four morphological features and one length based feature as described in the previous section. Thus our feature set is quite rich. Note that other than the usual features, semantic, position-based and user-based features are specific to the medical domain or to forum data.

### 6.3 Basic Results

First we considered word features, and learned a linear chain CRF model. We added other sets of features one by one, and observed variations in accuracy. Table 2 shows the accuracy in terms of precision, recall and F1. Note that these results are for an Order 1 linear-chain CRF. Accuracy is measured as ratio of the number of correct labelings of PE, MED and background to the total number of sentences in our dataset. Notice that the MED accuracy values are in general quite low compared to those of PE. As we will discuss later, accuracy is low for MED because our word-based features are not discriminative enough for the MED category.

From Table 2, we see that the accuracy keeps increasing as we add semantic UMLS based features, position based features and morphological features. However, length based features (word count), user-based faetures, and bigrams do not result in any improvements. We also tried trigrams, but did not observe any accuracy gains. Thus we find that semantic features and position-based features which are specific to the medical domain and the forum data respectively are helpful when added on top of word features, while generic features such as length-based features tend to not add value.

We also trained an order 2 CRF using the same

[5]http://www.healthboards.com
[6]http://mallet.cs.umass.edu/
[7]http://svmlight.joachims.org/

| Feature set | PE Prec | MED Prec | PE Recall | MED Recall | PE F1 | MED F1 | Accuracy % |
|---|---|---|---|---|---|---|---|
| Word | 0.60 | 0.49 | 0.65 | 0.36 | 0.62 | 0.42 | 63.43 |
| +Semantic | 0.61 | 0.52 | 0.68 | 0.37 | 0.64 | 0.43 | 65.05† |
| +Position | 0.63 | 0.54 | 0.7 | 0.34 | 0.66 | 0.42 | 65.45 |
| +Morphological | 0.64 | 0.52 | 0.69 | 0.36 | 0.66 | 0.42 | 65.70 |
| +WordCount | 0.62 | 0.51 | 0.70 | 0.33 | 0.66 | 0.40 | 65.23 |
| +Thread Creator | 0.62 | 0.51 | 0.71 | 0.34 | 0.66 | 0.41 | 65.49 |
| +Bigrams | 0.62 | 0.51 | 0.69 | 0.34 | 0.66 | 0.41 | 64.82 |

Table 2: Order 1 Linear Chain CRF. †Improvement over only word features significant at $0.05$-level, using Wilcoxon's signed-rank test

set of features. Results obtained were similar to order 1 CRFs and so we do not report them here. This shows that local dependencies are more important in medical forum data and global dependencies do not add further signal.

Further, we perform experiments using SVMs using the same set of features. Table 3 shows accuracy results on SVM. Again PE is detected with higher accuracy compared to MED. Unlike CRFs, SVMs do not incorporate the notion of local dependencies between sentences. However, we observe that SVMs outperform CRFs, as is evident from the results in Table 3. This is interesting, since it suggests that the SVM accuracy can potentially be further enhanced by incorporating such dependency information (e.g. in the form of new features). We leave this as part of future work.

Figure 2 shows an example of a forum post (which talks about allergy to dogs) being tagged using our CRF model.

<BKG>lori-lynn , </BKG>
<PE>you said he does well with the poms , but you also said he takes shots, so i wondered if the shots were for dog allergies</PE>
<PE>a lot of his friends have dogs , though , and he ' s so very allergic that he has trouble at their homes .</PE>
<MED>we opted not to go with the shots . </MED>
<BKG>i ' m still a little leary about adopting a dog . </BKG>
<BKG>i would just hate it if we did have reactions , because i know we ' d bond with the dog very quickly . </BKG>

Figure 2: Tagging example of a forum post

### 6.4 Feature selection

Incremental addition of different feature types did not lead to substantial improvement in performance. This suggests that none of the feature classes contains all "good" features. We therefore perform feature selection based on information gain and choose the top 4253 features from among all the features discussed earlier, based on a threshold for the gain. This results in improvement in the accuracy values over the previous best results (Table 4).

Among the word feature set, we found that important features were *allergy, alergies, food, hives, allergic, sinus, bread*. Among bigrams, *allergic_to, ear_infections, my_throat, are_allergic, to_gluten, food_allergies* have high information gain values. Among the UMLS based semantic groups, we found that *patf (Pathologic Function), dsyn (Disease or Syndrome), orch (Organic Chemical), phsu (Pharmacologic Substance), sosy (Sign or Symptom)* have high information gain values. Also looking at the word count feature, we notice that background sentences are generally short sentences. All these features are clearly highly discriminative.

### 6.5 Variation in training data size

We varied the amount of training data used for learning the models to observe the variation in performance with size of training data. Table 5 shows the variation in accuracy (PE F1, MED F1 and average accuracy) for different sizes of training data using CRFs. In general, we observe that accuracy improves as we increase the training data, but the degree varies with the feature sets used. We see similar trends in SVM also. These results show that it is possible to further improve prediction accuracy by obtaining additional training data.

### 6.6 Probing into the low MED accuracy

As observed in Tables 2 and 3, MED accuracy is quite low compared to PE accuracy. We wish to gain a deeper insight into why the MED accuracy suffers. Therefore, we plot the frequency of words in sentences marked as PE or MED versus the rank of the word as shown in the figure 3. We removed the stop words. Observe that for PE the

| Feature set | PE Prec | MED Prec | PE Recall | MED Recall | PE F1 | MED F1 | Accuracy % |
|---|---|---|---|---|---|---|---|
| Word | 0.65 | 0.52 | 0.71 | 0.28 | 0.68 | 0.36 | 66.13 |
| +Semantic | 0.73 | 0.54 | 0.73 | 0.38 | 0.73 | 0.45 | 71.02† |
| +Position | 0.71 | 0.52 | 0.71 | 0.35 | 0.71 | 0.42 | 69.61 |
| +Morphological | 0.72 | 0.53 | 0.72 | 0.38 | 0.72 | 0.44 | 70.28 |
| +WordCount | 0.74 | 0.54 | 0.72 | 0.37 | 0.73 | 0.44 | 71.55 |
| +Thread Creator | 0.74 | 0.56 | 0.72 | 0.39 | 0.73 | 0.46 | 72.02 |
| +Bigrams | 0.75 | 0.54 | 0.72 | 0.40 | 0.74 | 0.46 | 71.69 |

Table 3: SVM results. †Improvement over only word features significant at 0.05-level, using Wilcoxon's signed-rank test

| Classifier | PE Prec | PE Recall | PE F1 | MED Prec | MED Recall | MED F1 | Accuracy % |
|---|---|---|---|---|---|---|---|
| SVM (all* features) | 0.72 | 0.53 | 0.72 | 0.38 | 0.72 | 0.44 | 70.28 |
| SVM (selected features) | 0.75 | 0.75 | 0.75 | 0.61 | 0.33 | 0.44 | 75.08† |
| CRF (all* features) | 0.64 | 0.52 | 0.69 | 0.36 | 0.66 | 0.42 | 65.70 |
| CRF (selected features) | 0.60 | 0.77 | 0.67 | 0.58 | 0.37 | 0.45 | 65.93† |

Table 4: Accuracy using the best feature set. (*Word +Semantic +Position +Morphological features). †Improvement over all* features significant at 0.05-level, using Wilcoxon's signed-rank test

curve is quite steep. This indicates that there are some discriminative words which have very high frequency and so the word features observed in the training set also get fired for sentences in the test set with high probability. While for MED, we observe that most of the words have very low frequencies. This basically means that discriminative words for MED may not occur with good enough frequency. So, many of the word features that show up in the training set may not appear in the test data. Hence, MED accuracy suffers.

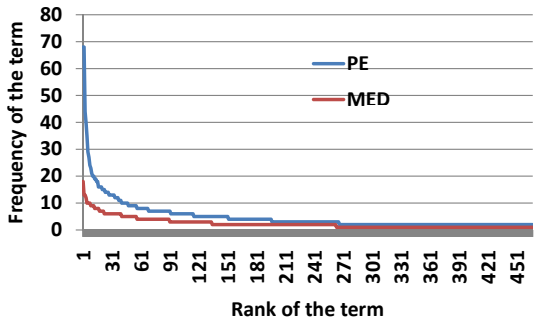


Figure 3: Freq of words vs rank for PE and MED

### 6.7 Multi-class vs Single class categorization

Note that our task is quite different from plain sentence categorization task. We observe that there is a dependence between the categories (PE/MED) that we are trying to predict per sentence. For example, considering 100% training data, Table 6 compares the precision, recall and F1 values when SVM and CRF are trained as single class classifiers using word+semantic features with the multi-class results obtained previously. Results are generally better when we do multi-class categorization versus single-class categorization. This trend was reflected for other featuresets also.

| | PE | MED | Backgnd | EOP |
|---|---|---|---|---|
| PE | 0.54 | 0.13 | 0.28 | 0.05 |
| MED | 0.15 | 0.51 | 0.30 | 0.04 |
| Backgnd | 0.18 | 0.08 | 0.54 | 0.20 |
| BOP | 0.40 | 0.07 | 0.53 | 0.0 |

Table 7: Transition probability values

### 6.8 Analysis of transition probabilities

Table 7 shows the transition probabilities from one category to another as calculated based on our labelled dataset. BOP is beginning of posting and EOP is end of posting. Note that posts often start with a PE or a background sentence and often end with a background sentence. Also, consecutive sentences within a posting tend to belong to the same category.

### 6.9 Error analysis

We also perform some error analysis on results using the best feature set. Table 8 shows the confusion matrix for CRF/SVM. We observe many of the MED errors are because an MED sentence

| Feature set | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| Word | 0.59/0.21/0.57 | 0.6/0.36/0.60 | 0.61/0.39/0.62 | 0.62/0.42/0.63 |
| +Semantic | 0.61/0.17/0.59 | 0.63/0.32/0.61 | 0.64/0.38/0.63 | 0.64/0.43/0.65 |
| +Position | 0.59/0.18/0.56 | 0.64/0.29/0.60 | 0.65/0.33/0.62 | 0.66/0.42/0.65 |
| +Morphological | 0.6/0.19/0.57 | 0.64/0.32/0.61 | 0.65/0.37/0.63 | 0.66/0.42/0.65 |
| Best | 0.61/0.18/0.65 | 0.66/0.28/0.64 | 0.66/0.38/0.66 | 0.69/0.43/0.68 |

Table 5: Precision, recall, and F value for various sizes of training data set.

| Classifier Type | PE Prec | PE Recall | PE F1 | MED Prec | MED Recall | MED F1 |
|---|---|---|---|---|---|---|
| SVM PE vs BKG | 0.79 | 0.64 | 0.71 | - | - | - |
| SVM MED vs BKG | - | - | - | 0.6 | 0.28 | 0.39 |
| SVM Multi-class | 0.73 | 0.73 | 0.73 | 0.54 | 0.38 | 0.45 |
| CRF PE vs BKG | 0.68 | 0.64 | 0.66 | - | - | - |
| CRF MED vs BKG | - | - | - | 0.53 | 0.3 | 0.39 |
| CRF Multi-class | 0.61 | 0.68 | 0.64 | 0.52 | 0.37 | 0.43 |

Table 6: Multi-class vs Single-class categorization with word+semantic features

| | PE | MED | Backgnd |
|---|---|---|---|
| PE | 424/404 | 37/37 | 81/101 |
| MED | 102/70 | 107/95 | 81/125 |
| Backgnd | 164/62 | 55/21 | 618/754 |

Table 8: Confusion matrix showing counts of actual vs predicted labels for (Best CRF Classifier/Best SVM Classifier)

often gets marked as PE. This basically happens because some sentences contain both PE and MED. Other than that some of the PE keywords are also present in MED sentences, and since the few discriminative MED keywords are quite low in frequency, MED accuracy suffers. E.g. The sentence *"i'm still on antibiotics for the infection but they don't seem to be doing any good anymore."* was labeled as MED but marked as PE by the CRF. The sentence clearly talks about a medication. However, the keyword *"infection"* is often observed in PE sentences and so the CRF marks the sentence as PE.

## 7 Conclusion

In this paper, we studied a novel shallow information extraction task where the goal is to extract relevant sentences to a predefined set of categories that describe a medical case. We proposed to solve the problem using supervised learning and explored two representative approaches (i.e., CRF and SVM). We proposed and studied two different types of novel features for this task, including generalized terms and forum structure features. We also created the first test set for evaluating this problem. Our experiment results show that (1) the proposed new features are effective for improving the extraction accuracy, and (2) it is feasible to automatically extract medical cases in this way, with the best prediction accuracy above 75%.

Our work can be further extended in several ways. First, since constructing a test set is labor-intensive, we could only afford experimenting with a relatively small data set. It would be interesting to further test the proposed features on larger data set. Second, while in CRF, we have shown adding dependency features improves performance, it is unclear how to evaluate this potential benefit with SVM. Since SVM generally outperforms CRF for this task, it would be very interesting to further explore how we can extend SVM to incorporate dependency.

## 8 Acknowledgement

We thank the anonymous reviewers for their useful comments. This paper is based upon work supported in part by an IBM Faculty Award, an Alfred P. Sloan Research Fellowship, an AFOSR MURI Grant FA9550-08-1-0265, and by the National Science Foundation under grants IIS-0347933, IIS-0713581, IIS-0713571, and CNS-0834709.

## References

Aone, Chinatsu and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *ANLP*.

Boser, Bernhard E., Isabelle Guyon, and Vladimir Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152.

Buttler, David, Ling Liu, and Calton Pu. 2001. A fully automated object extraction system for the world wide web. In *ICDCS*.

Cios, Krzysztof J. and William Moore. 2002. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26:1–24.

Cong, Gao, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 467–474, New York, NY, USA. ACM.

Cui, Hang, Vibhu Mittal, and Mayur Datar. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proc. of the National Conf. on Artificial Intelligence*, pages 1265–1270.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proc. of WWW*, pages 519–528.

Feng, Donghui, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An intelligent discussion-bot for answering student queries in threaded discussions. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 171–177, New York, NY, USA. ACM.

Harabagiu, Sanda and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912, Morristown, NJ, USA. Association for Computational Linguistics.

Heinze, Daniel T., Mark L. Morsch, and John Holbrook. 2002. Mining free-text medical records. In *Proceedings of the AMIA Annual Symposium*.

Holt, Alec, Isabelle Bichindaritz, Rainer Schmidt, and Petra Perner. 2005. Medical applications in case-based reasoning. *Knowl. Eng. Rev.*, 20(3):289–292.

Kim, Hyun Duk and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *CIKM*, pages 385–394.

Laender, Alberto H. F., Berthier A. Ribeiro-neto, Altigran S. da Silva, and Juliana S. Teixeira. 2002. A brief survey of web data extraction tools. *SIGMOD Record*.

Lazarus, R, K P Kleinman, I Dashevsky, A DeMaria, and R Platt. 2001. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health*, 1:9.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment Classification using Machine Learning techniques. In *Proc. of EMNLP*, pages 79–86.

Prabowo, Rudy and Mike Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, April.

Sarawagi, Sunita. 2008. Information extraction. *Foundations and Trends in Databases*, 1.

Sutton, Charles and Andrew Mccallum, 2006. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press.

Zhou, Xiaohua, Hyoil Han, Isaac Chankai, Ann Prestrud, and Ari Brooks. 2006. Approaches to text mining for clinical medical records. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 235–239, New York, NY, USA. ACM.