Extracting the Meaning of Medical Concept Correlations

Meliha Yetisgen-Yildiz
The Information School, University of Washington, Seattle, USA
melihay@u.washington.edu

Wanda Pratt
The Information School, Biomedical and Health Informatics, University of Washington, Seattle, USA
wpratt@u.washington.edu

ABSTRACT
In this paper, we propose a new method to extract the meaning of medical concept correlations from MEDLINE abstract sentences. Our method incorporates a medical knowledge base, natural language processing approaches, and text classification methods. We describe how we automatically created the training sets and report the results of our initial experiments.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Medical Information Systems.

General Terms
Algorithms.

Keywords
Text classification, knowledge-based systems.

INTRODUCTION
Correlation identification methods based on medical concept co-occurrences (i.e. mutual information) have been commonly used in medical text mining systems. Medical concepts co-occur together for many different reasons. For example, possible explanations of a correlation between a disease or symptom, *S*, and a chemical or drug, *D*, can be listed as; (1) *D* is used to treat *S*, (2) a side affect of *D* causes *S*, or (3) *D* prevents *S*. Our ultimate aim is to propose a method to extract the meaning of the correlation between two medical concepts from the sentences that the concepts co-occur. In this paper, we limit it to correlations between a disease or symptom and a drug or chemical.

METHOD
In the following sections, we will describe the main components of the method we designed to extract the meaning of medical concept correlations.

Extraction of Medical Concepts
We use the Unified Medical Language Systems (UMLS) as the main knowledge source [4] to extract the medical concepts. National Library of Medicine (NLM) created this publicly available medical knowledge base by unifying hundreds of medical knowledge bases. To identify the medical concepts in MEDLINE abstracts, we use MMTx, an NLP library created by NLM [3]. We use the functions available in MMTx to break the abstracts into sentences and to map the sentences to the UMLS concepts. We store the sentences along with the extracted UMLS concepts in a sentence database.

Extraction of Relationships
We use UMLS Semantic Network to identify the list of potential relationships among correlated medical concepts. Semantic Network is a directed graph composed of 135 categories called *semantic types* and 49 relations defined between the semantic types. Each medical concept in UMLS is mapped to at least one semantic type. To decrease the size of the network, we grouped the semantic types under *semantic groups* and created a semantic group graph. In UMLS, there are 15 semantic groups. In our extraction approach, we use the semantic group graph as a guide to identify the meaning of medical concept correlations. We first retrieve the semantic groups of the medical concepts from the UMLS and extract the relations between the semantic groups from the graph. Suppose we want to extract the relations between *ergotamine* and *migraine*. The semantic group of *ergotamine* is *Chemicals and Drugs* and the semantic group of *migraine* is *Disorders.* In the semantic group graph, *Chemicals and Drugs* is connected to *Disorders* through seven different relations, *affects, causes, complicates, diagnoses, prevents, indicates,* and *treats*. We use the identified relations as the list of possible meanings of the correlation. However, some of relations are contradictory to each other (e.g., *causes* and *treats*) and the challenge is to select the correct relations for the given medical concepts. Deciding which relations hold for a given correlation between two medical concepts can be seen as a classification problem. For a given correlation between the concepts t_1 and t_2, suppose s_1 is the semantic group of t_1, s_2 is the semantic group of t_2, R is the set of relations between s_1 and s_2, and D is the set of sentences that include both t_1 and t_2. The classification problem is to label the sentences in D with the relations from R. To accomplish this, we implemented a Naïve-Bayes classifier [2].

K-CAP'07, October 28–31, 2007, Whistler, British Columbia, Canada.

Creating the Training Sets for Classifiers

There are a number of manually annotated medical corpora that provide information about proteins and their interactions (e.g., GENIA, GENETAG, and PennBioIE). However, the information contained in those corpora is focused on protein interactions and is not sufficient to capture the type of knowledge we want with our information extraction task. To overcome the expensive manual labeling process while creating the training sets for the classifiers, we use a semi-automated labeling approach. In the semantic group graph, a link between two semantic groups, s_i and s_j, is represented as (*semantic group* s_i – *relation* r_k – *semantic group* s_j) triple. Our objective is to train one binary classifier for each link available in the semantic group graph but we do not have any training set. To create the training sets, for each ($s_i - r_k - s_j$) triple, we first manually identify medical concept couples, (c_1, c_2), where the semantic group of c_1 is s_i, the semantic group of c_2 is s_j, and the relation between c_1 and c_2 is known to be r_k. We call c_1 and c_2 as *seed concept couple*. As a next step, we query the sentence database to extract sentences that include both the selected concepts, c_1 and c_2, and label those sentences as positive examples in the training set for the ($s_i - r_k - s_j$) triple. However, to train a binary classifier, we also need negative examples. We use a labeling heuristic called PNLH (**P**ositive Examples and **N**egative Examples **L**abeling **H**euristic) to identify the negative examples [1]. PNLH was proposed for situations similar to ours where there is a small set of positive examples, no negative examples, and a large set of unlabeled examples that includes both positive and negative examples. We randomly select a set of sentences from sentence database that include medical concepts with semantic types s_i and s_j, and mark those sentences as unlabeled examples. We then apply PNLH's first iteration to identify the negative examples in the set of unlabeled examples.

Elimination of Descriptive Clinical Trial Setting (DCTS) Sentences

While investigating the characteristics of sentences in MEDLINE abstracts, we noticed that many sentences describe only the experiment settings of clinical trial studies without providing any information about the correlations between the medical concepts (i.e. "In this randomized, double-blind, parallel-group phase-II study, 40 patients with acute migraine attacks alternately received iVPA 800 mg or iLAS 1000 mg."). To eliminate such descriptive sentences, we randomly selected 220 sentences, manually labeled them, and trained a naïve-bayes classifier to eliminate the DCTS sentences.

EXPERIMENTS & RESULTS

In this section, we present our initial performance results of the classifier trained for (*Chemicals & Drugs – treats – Disorders*) triple. In our experiments, we used a portion of MEDLINE composed of 397,909 abstracts published in 2005. We created a sentence database composed of 1,777,829 sentences from those abstracts.

To identify the positive examples, we used the top 10 most sold US drugs in 2006 (www.drugs.com) and the corresponding target diseases as the seed concept couples. There were 172 sentences in the sentence database that included the seed concepts. 80 of those sentences were classified as non-DCTS sentences and marked as positive examples. To identify the negative examples, we randomly selected 1000 sentences from the sentence database that included medical concept couples with semantic groups *Chemicals & Drugs* and *Disorders*, and marked those sentences as unlabeled examples. 721 out of 1000 sentences were classified as non-DCTS and 172 out of the 721 sentences were selected as negative examples by PNLH. We trained the classifier with the selected 80 positive and 172 negative training examples.

To create a test set, we queried the sentence database for sentences that are about *Migraine* and any medical concept with the semantic group *Chemicals & Drugs* and marked the retrieved 204 sentences as test sentences. 119 of the test sentences were non-DCTS sentences and our classifier classified 72 out of 119 sentences as treatment sentences. To create a gold standard for the evaluation, we manually checked all 119 sentences and labeled the ones that are about a migraine treatment with a chemical or drug. The classifier produced 74% precision and 89% recall.

CONCLUSION

In this paper, we described a new semi-automated method to extract the meaning of medical term correlations from MEDLINE abstracts. We have not evaluated each step of our method in detail yet. However, based on the good performance results reported for one type of relation, our general extraction approach to generate positive and negative training examples is promising.

ACKNOWLEDGMENTS

The National Science Foundation, award IIS-0133973, funded this work.

REFERENCES

[1] Fung, G.P.C., et al.., Text Classification without Negative Examples Revisit, IEEE Transactions on Knowledge and Data Engineering 18 (2006), pp. 6-20.

[2] Mitchell, T.M., Machine Learning, McGraw-Hill, (1997).

[3] NLM, MMTx Project, Available at: http://mmtx.nlm.nih.gov/docs.shtml (2006).

[4] NLM, UMLS Fact Sheet, Available at: http://www.nlm.nih.gov/pubs/factsheets/umls.html (2006).