Finding the Meaning of Medical Concept Correlations

Meliha Yetisgen-Yildiz, Ph.D.,[1] Wanda Pratt, Ph.D.[1,2]

[1] The Information School, University of Washington, Seattle, USA.
[2] Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, USA.

Abstract

Correlation identification methods based on concept co-occurrences have been commonly used on medical free texts. However, concepts co-occur for different reasons, and generalizable approaches to determine the meaning of those co-occurrences are needed. In this paper, we propose a new extraction approach that incorporates UMLS and text classification methods to identify the semantics of the relationships between co-occurring concepts in MEDLINE abstracts. The major difficulty of our approach is the lack of annotated sentences for training and testing purposes. We describe how we semi-automatically annotate the sentences with a combination of heuristics and a partially supervised classification method. In our evaluations, we focus on extracting the meaning of only the correlations between drugs or chemicals and disorders, and we limit the meaning to treats and causes. Based on the good performance results, we believe that our approach shows great promise for tackling the difficult relationship-identification problem in medical free text.

Introduction

A vast amount of medical knowledge is locked up in free-text documents. Current research to extract semantics from these sources focuses on identifying the medical concepts (e.g. genes, proteins, diseases) from free text and extracting the correlations between the identified concepts (e.g. protein-protein interactions)[1]. Various sophisticated natural language processing (NLP), statistical, and machine learning approaches have been developed to identify phrases that refer to medical concepts in text. Many of the correlation extraction approaches are based on statistical analysis of term co-occurrences in text, but these approaches do not distinguish among the many ways in which medical concepts co-occur. For example, possible explanations for 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. To be able to identify which of the listed meanings hold for a given disease-drug correlation would be very helpful for the users of the text mining systems. In this paper, we propose a new extraction approach to identify the meaning of correlated medical concepts from MEDLINE abstract sentences. In our approach, we use the UMLS Semantic Network[2] to extract the list of potential meanings for a given correlation between two medical concepts and use text classification approaches to identify which of the extracted meanings are true. One major challenge for this approach is the lack of annotated training corpora available for the scale of this extraction. To overcome the expensive manual annotation process, we manually identify seed examples to select the positive sentences and use a partially supervised classification method that is based on the Expectation-Maximization (EM) algorithm[3] and the Naïve-Bayes classification[4] to select the negative examples.

Inferring relations between the co-occurring medical concepts based on the UMLS Semantic Network has been the subject of previous research[5,6]. Burgun and Bodenreider have used the relations defined in the Semantic Network to infer the possible meaning of the co-occurring concepts in MEDLINE[5]. They have shown the effectiveness of their approach with a detailed statistical analysis. However, they have not proposed any methods to disambiguate between conflicting relations, such as *treats* and *causes* defined between the semantic groups *disorders* and *chemicals & drugs*. In this paper, we attempt to solve this disambiguation problem by using machine learning methods.

Ahlers et.al. have described an NLP system (Enhanced SemRep) that is based on the domain knowledge available in UMLS Semantic Network[6]. The output of their system is in the form of semantic predictions that represent assertions from the MEDLINE abstracts expressing a range of specific relations in pharmacogenomics. Our extraction approach is more general than those previously published approaches because it can be applied to any type of medical concepts, rather than limiting it to specific domains such as protein-protein or chemical-disease interactions. Also, in this paper, we apply our method to only MEDLINE abstracts. However, it may be applicable to other knowledge sources such as patient records and web documents.

Methods

In this section, we will describe the main components of our extraction approach and semi-automated annotation method.

Identification of the Medical Concepts from MEDLINE Abstracts

We use the UMLS as the main knowledge source to extract the medical concepts from medical text[2]. To identify the medical concepts in MEDLINE abstracts, we use MMTx, an NLP library created by NLM[7]. We use the functions available in MMTx to break the abstracts into sentences and to map the sentences to the UMLS Metathesaurus phrases. We also use the UMLS Metathesaurus concept unique identifiers (CUI) to combine synonymous UMLS phrases. As an example, the following synonymous medical phrases; *heart arrest*, *cardiac arrest*, and *ventricular asystolia* are grouped under a unique CUI: C0018790. We give each sentence a unique identifier that is a combination of the MEDLINE document identifier (PMID) and the order of the sentence in the abstract and store it with the extracted UMLS concept identifiers in a sentence database. We use this database to search for sentences that include the given correlated medical concepts.

Identification of the Relationships

We use the UMLS Semantic Network to identify the list of potential relationships among correlated medical concepts. To decrease the size of the semantic network, we grouped the semantic types under the *semantic groups*[8] and created a semantic group graph. In UMLS, there are 15 semantic groups to categorize 135 semantic types. For example, the semantic type of the medical concept *migraine* is *Disease and Syndrome* and the semantic type of the medical concept *panic disorder* is *Mental or Behavioral Dysfunction*. The semantic group of both semantic types is *Disorders*. A portion of the semantic group graph is represented in Figure 1.

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 six different relations, *affects*, *causes*, *complicates*, *diagnoses, prevents,* 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. We have posed this challenge of deciding which relations hold for a given correlation between two medical concepts as a classification problem.

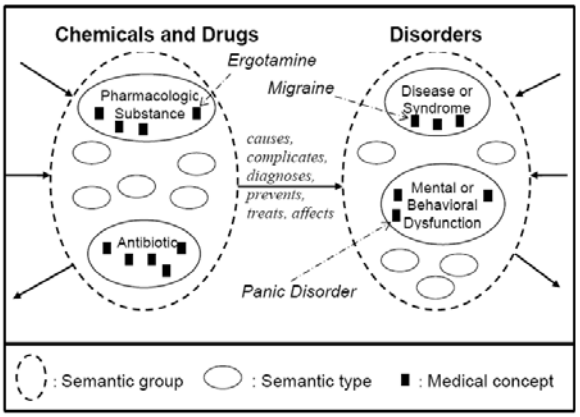

Figure 1. An example of how semantic groups, semantic types, biomedical concepts and relations are represented in the semantic group graph.

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.

Creation of 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)[1]. Because our ultimate aim is to identify the meaning of correlations between any type of medical terms, rather than just those between proteins, those available corpora are 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 annotated 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

$(\textit{semantic group}\ s_i - \textit{relation}\ r_k - \textit{semantic group}\ s_j)$

triple. Our objective is to train one binary classifier for each link in the semantic group graph, but we do not have any annotated sentences that can be used as training sets.

To annotate sentences 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 the *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 annotate those sentences as positive sentences for the ($s_i - r_k - s_j$) triple.

A good seed concept couple example that can be used to identify positive sentences for the classifier of (*Chemicals and Drugs – treats – Disorders*) triple, would be (Chemical and Drug: *Sumatriptan*, Disorder: *Migraine*) because *sumatriptan* is a 5-HT1 receptor agonist which is widely used in the treatment of *migraines*. An example sentence that includes both *sumatriptan* and *migraine* is "*Oral sumatriptan (100 mg) is an effective and well tolerated acute treatment for patients who report menstrually related migraine*." The common characteristic of such sentences is that they include information about using *sumatriptan* in the treatment of *migraine* or the potential benefits of *sumatriptan* in relieving *migraine* related symptoms. In our approach of selecting the seed concept couples, we hypothesize that the sentences for other chemicals used in the treatment of other diseases will be contextually similar.

We use the seed couples to identify the positive sentences, but to train a classifier, we also need negative examples. We use a technique based on a combination of Expectation-Maximization (EM) with a Naïve-Bayes classifier to identify the negative examples[9-10]. This technique was proposed for situations similar to ours where there is a small set of positive examples, none[9] or a small set[10] of negative examples, and a large set of unlabeled examples that includes both positive and negative examples. We randomly select a set of sentences from the sentence database that include medical concepts with semantic groups s_i and s_j, and mark those sentences as unlabeled examples. We then apply EM to identify the negative examples in the set of unlabeled examples by using the set of positive examples. The details of Naïve-Bayes text classification and EM will be presented in the following sections.

Naïve-Bayes Text Classification

Naïve-Bayes is one of the widely used techniques for text classification[4]. The learning task for the Naïve-Bayes classifier is to use a set of training sentences to estimate the model parameters then use the estimated model to classify the new sentences.

We used *bag-of-words* approach to represent the sentences. We converted the sentences into lower-case and divided them into words by using white spaces. To decrease the number of features, we stemmed the words by using *Porter stemmer*[11] and grouped all numerical values under a single label. Suppose we have a set of training sentences *D* and each sentence $d_i \in D$ is assigned to a category $c_j \in C$. The purpose of a Naïve-Bayes classifier to label a given test sentence t_i with a category $c_j \in C$ that produces highest $P(c_j \mid t_i)$ given the set of training sentences, *D*. $P(c_j \mid t_i)$ is calculated with the following formula:

$$P(c_j \mid t_i) = \frac{P(c_j)\prod_{k=1}^{|t_i|} P(w_k \mid c_j)}{\sum_{m=1}^{|C|} P(c_m)\prod_{k=1}^{|t_i|} P(w_k \mid c_m)} \quad (1)$$

where $P(c_j)$ is the class probability of c_j, w_k is a word that appears in t_i's representation, and $P(w_k \mid c_j)$ is the probability of w_k given c_j. $P(c_j)$ and $P(w_k \mid c_j)$ are calculated with the following formulas:

$$P(c_j) = \frac{\sum_{i=1}^{|D|} P(c_j \mid d_i)}{|D|} \quad (2)$$

$$P(w_k \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_k, d_i) P(c_j \mid d_i)}{|V| + \sum_{l=1}^{|V|}\sum_{i=1}^{|D|} N(w_l, d_i) P(c_j \mid d_i)} \quad (3)$$

where $N(w_k, d_i)$ is the number of times word w_k appears in training sentence d_i and *V* is the set of distinct words that appear in the complete set of training sentences.

Identification of Negative Training Sentences with Expectation-Maximization Algorithm (EM)

The expectation-maximization (EM) algorithm is a general framework for estimating the parameters of a probability model when the data has missing values[3]. We applied the algorithm to identify the missing labels of the drug-disease sentences. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the

parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated.

The EM algorithm can be applied to identify negative sentences as follows. Suppose *P* is the set of positive sentences and *U* is the set of unlabeled sentences. We label each positive sentence p_i in *P* with c_+ (positive class label) and each unlabeled sentence u_j in *U* with c_- (negative class label). With this labeling, $(P(c_+ \mid p_i) = 1, P(c_- \mid p_i) = 0)$ and $(P(c_+ \mid u_j) = 0, P(c_- \mid u_j) = 1)$. We use this initial labeling to train a Naïve-Bayes classifier. This classifier is used to classify the unlabeled sentences in *U*. Each u_j in *U* is assigned to a probabilistic class label. After each $P(c_+ \mid u_j)$ is updated, a new classifier is trained on unlabeled sentences in *U* with the new class probabilities and the positive sentences in *P*. This iterative process continues until convergence.

We use the final probabilistic class labels of the sentences in *U* to identify the negative sentences. A key point of the approach is that we leave the class probability of each positive sentence p_i in *P* unchanged throughout the process.

Evaluation

In this section, we present our initial performance results of the two classifiers trained for (*Chemicals & Drugs – treats – Disorders*) and (*Chemicals & Drugs – causes – Disorders*) triples. 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.

Selection of the Seed Concept Couples

To identify the positive examples, we used information about the top 20 most sold US drugs in 2006 listed on www.drugs.com website. By using the drug descriptions available on the website and UMLS as the main knowledge sources, for each selected drug, we identified the corresponding target diseases and used them as the seed concept couples to extract the positive examples for the *treats* classifier (e.g. Drug: *seroquel* – Disease: *schizophrenia*). In the same way, for each drug, we identified the potential side-effects and used them as seed concept couples to extract the positive examples for the *causes* classifier (e.g. Drug: *seroquel* – Side-effect: *weight gain)*. Because we represented the content of the abstracts with UMLS concepts, we checked the selected diseases and side-effects against UMLS and eliminated the ones that were not in the UMLS.

From the descriptions of the selected 20 drugs, we identified 85 seed couples for *treats* classifier and 202 seed couples for *causes* classifier. We queried the sentence database for the seed couples and extracted 496 positive sentences for treats classifier and 63 positive sentences for the *causes* classifier.

Elimination of the Descriptive Experiment Setting (DES) Sentences

While investigating the characteristics of positive sentences identified for *treats* and *causes* classifiers, we noticed that many sentences describe only the experiment settings (e.g. "*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.* ") without providing any information about the correlations between the drugs or diseases. To eliminate such setting sentences, we first selected 300 sentences about drugs and diseases from the sentence database, manually labeled them as DES or non-DES and used those labeled sentences to train a Naïve-Bayes classifier. To estimate the performance of our classifier, we applied 5-fold cross validation. The Naïve-Bayes classifier produced on the average 92% precision and 91% recall in identifying the non-DES sentences. We thought these results were sufficient to justify the semi-automated sentence elimination approach.

Selection of the Positive, Unlabeled, and Negative Sentences

As described in the previous sections, we created the positive sentence sets by querying the sentence database for the seed couples. We created the unlabeled sentence set by randomly selecting from the sentence database 1000 sentences that included medical concept couples with semantic groups *Chemicals & Drugs* and *Disorders*. There were no overlaps between the positive and the unlabeled sentence sets. We then filtered the DES sentences from both the positive and unlabeled sentence sets by using the Naïve-Bayes classifier described in the previous section. To extract the negative sentences for each classifier, we ran EM on the positive and unlabeled sentence sets composed of only non-DES sentences. Table 1 includes the statistical information about the positive, unlabeled and negative sentence sets.

We manually assessed the quality of the positive and negative sentences. We first randomly selected 100 sentences from the union of non-DES positive

sentence sets and 100 sentences from the union of non-DES negative sentence sets created for *treats* and *causes* classifiers. We then manually checked the correctness of the sentences and calculated precision. 93% of the positive sentences included an indication of either a treatment or side-effect. 95% of the negative sentences were neither about a treatment nor side-effect.

	Positive		Unlabeled		Negative
	Mixed	Non-DES	Mixed	Non-DES	Non-DES
Treats	496	367	1000	887	693
Causes	63	60	1000	887	839

Table 1. Statistical information about positive, unlabeled, and negative sets for *treats* and *causes* classifiers.

Calculation of Performance Metrics

To measure the performance of the classifiers, we used 5-fold cross validation on selected non-DES positive and non-DES negative sentences. The performance results for *treats* and *causes* classifiers are presented in Table 2.

	Precision	Recall	Accuracy
Treats	97%	98%	98%
Causes	86%	46%	95%

Table 2. 5-fold cross validation results for treats and causes classifiers.

As can be seen from the table, when trained with 80% of the annotated sentences, the t*reats* classifier successfully identified the sentences that indicated a potential treatment of a disorder with a chemical or drug from the remaining 20% of the annotated sentences. On the other hand, the performance of the *causes* classifier was lower than that of *treats* classifier. This performance difference can be explained with the size differences between the positive sentence sets of the two classifiers. Although there were many seed concept couples identified for the *causes* classifier, there were only 60 non-DES positive sentences, which was clearly insufficient to capture the characteristics of the sentences that indicated a disorder caused by a chemical or drug.

Conclusion

There are two main contributions of this paper. The first contribution is our new semi-automated extraction approach to identify the meaning of medical term correlations from MEDLINE abstract sentences. In our extraction approach, we combined the knowledge available in the UMLS Metathesaurus and Semantic Network with the text classification methods to find the correct meanings of a given correlation. The second contribution is our semi-automated way of annotating the sentences to create the training sets needed in our extraction approach.

In this paper, we presented our initial evaluation results for the two types of relations, *treats* and *causes* between drugs and diseases. More studies are needed to evaluate each step of our method in detail and to evaluate it on a larger selection of relations. However, based on the good performance results reported, we believe that our general extraction approach and semi-automated annotation method are promising to extract the meanings of a variety of medical concept co-correlations.

References

1. Ananiadou, S., and Mcnaught, J.: Text Mining for Biology and Biomedicine: Artech House Publishers; 2005.
2. NLM: UMLS Fact Sheet. Available at: http://wwwnlmnihgov/pubs/factsheets/umlshtml.
3. Dempster A.P., Laird, N.M., and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 1977, B 39:1-38.
4. Mitchell, T.: Machine Learning: McGraw Hill; 1997.
5. Burgun, A., Bodenreider, O.: Methods for exploring the semantics of the relationships between co-occuring UMLS concepts. In: Proceedings of MEDINFO; 2001.
6. Ahlers, C.B., Fiszman, M., Demner-Fushman, D., Lang, F-M, and Rindflesch, T.C.: Extracting Semantic Predictions from MEDLINE Citations for Pharmacogenomics. In: Proceedings of PSB; 2007.
7. NLM: MetaMap. Available at: http://mmtxnlmnihgov/docsshtml.
8. McCray, A.T., Burgun, A., and Bodenreider, O.: Aggregating UMLS Semantic Types for Reducing Conceptual Complexity, In: Proceedings of MEDINFO; 2001.
9. Liu, B., Lee, W.S., Yu, P.S., and Li, X.: Partially Supervised Classification of Text Documents. In: Proceedings of ICML; 2002.
10. Nigam, K., McCallum, A., Thrun, S., and Mitchell, T.: Learning to Classify Text from Labeled and Unlabeled Documents. In: Proceedings of AAAI; 1998.
11. Porter M.F.: An algorithm for suffix stripping. Program 1980, 14(3):130−137.