6.864 Final Project
Classification of Research Article Findings
based on Cochrane Medical Reviews

Battushig Myanganbayar
Instructors: Prof. Regina Barzilay, Franck Dernoncourt

December 14, 2015
Massachusetts Institute of Technology

Abstract

In medical community, conclusively determining effectiveness of treatment or medicine for certain symptoms requires careful review of all related papers, and meticulous analysis on statistical significance of the facts stated on those papers. This entire procedure takes substantial amount of time, and resources.

To make process much more efficient, and faster, we designed natural language processing algorithm that analyzes all the related articles, and able to make conclusion about given hypothesis. If majority of referenced articles supports the argument, then final decision for given hypothesis would be Yes, and No if otherwise.

Medical articles are complex in semantics, and summarizing them from raw text requires a lot of supervised data. To overcome such challenges, and still be able to test the NLP algorithm of our interest, we will use Cochrane Reviews article, and PubMed citations referenced in particular article.

In this project, Cochrane Reviews article used to determine the hypothesis, as well as underlying ground truth. Referenced PubMed articles, then, served as a input to NLP, and predicted decision is checked against ground truth to algorithm. We found promising results on accuracy of NLP algorithm[1].

[1]You can view implementation code on https://github.com/Tushigee/CochraneProj.git

Contents

List of Figures

List of Tables

1 Objective

The long-term vision of this project is to choose and evaluate most relevant papers, or articles from Internet to decide whether evidence supports or refutes the hypothesis such as whether vitamin C is good for cough.

However, determining most relevant evidence from Internet is very hard task, and this part will be addressed for the future. We will now focus on decision algorithm, and approaches improving accuracy of evaluating reference article.

In this report, we will hypothesize, implement, and test the algorithm that is able to make correct binary decision about claim based on polarity of reference articles.

2 Dataset

It is important for application that ground truth of hypothesis is accurate, and all referenced articles are highly relevant to the hypothesis. Therefore, we are using articles, and references from Cochrane Review, a medical database.

Cochrane Review has collective summary of about 6,000 articles, and considered one of the most credible reviews in the medical community. Each Cochrane Review Article has following information Objective, Main Results, and Authors' conclusions included on it.

For each Cochrane article, there is a list of referenced articles that is hosted on different websites such as PubMed, Crossref, and CAS. Reference articles on PubMed has brief summary of each article. On the other hand, Crossref and CAS only have PDF version to download. Therefore, references from PubMed are used in this project.

Following is an example of article from Cochrane Review:

> OBJECTIVES: *To evaluate the impact of cancer genetic risk-assessment services on patients at risk of familial breast cancer.*
>
> MAIN RESULTS: *In this review update, we included five new trials, bringing the total number of included studies to eight. The included trials (pertaining to 10 papers), provided data on 1973 participants and assessed the impact of cancer genetic risk assessment on outcomes including perceived risk of inherited cancer, and psychological distress. This review suggests that cancer genetic risk-assessment services help to reduce distress, improve the accuracy of the perceived risk of breast cancer, and increase knowledge about breast cancer and genetics.*
>
> AUTHORS CONCLUSION: ***This review found favorable outcomes for patients*** *after risk assessment for familial breast cancer. However, there were too few papers to make any significant conclusions about how best to deliver cancer genetic risk-assessment services. Further research is needed assessing the best means of delivering cancer risk assessment.*

As we can see from example above, this review should be classified as YES by our prediction algorithm since Authors Conclusion states positive comments

about the Objective.

Following is an example reference article cited by Cochrane article mentioned above:

> RESULTS: *Although statistically significantly greater improvement in knowledge about breast cancer was found in the trial group (P: =.05), differences between groups in other psychological outcomes were not statistically significant.* ***Women in both groups experienced statistically significant reductions in anxiety and found attending the clinics to be highly satisfying****. An initial specialist genetic assessment cost pound 14.27 (U.S. $22.55) more than a consultation with a breast surgeon.*
>
> CONCLUSIONS: ***There may be benefit in providing specialist genetics services*** *to all women with a family history of breast cancer. Further investigation of factors that may mediate the impact of genetic assessment is in progress and may reveal subgroups of women who would benefit from specialist genetics services.*

We can also see that this reference articles finding was positive. Therefore, by analyzing polarity of reference articles from PubMed, we are able to reasonably well predict the conclusion being YES or NO.

2.1 Data Scraping

Cochrane database uses dynamic JavaScript puzzle to send articles to client. Therefore, significant amount of effort is dedicated to download total of 6,325 articles from Cochrane Review. Browser fingerprint based rate control also precluded scraping algorithm to download articles at faster rate.

With limitation imposed above, we used Selenium Web testing framework to go around rate control limitation, as well as XPath based parser to deal with dynamically changing structure of HTML page. Total of 4 days spent to scrape entire Cochrane Reviews Database.

References hosted on PubMed, on the other hand, have no restrictions on scraping, and simple http library used to download total of 47,110 reference articles.

Following is the block diagram summary of data scraping process

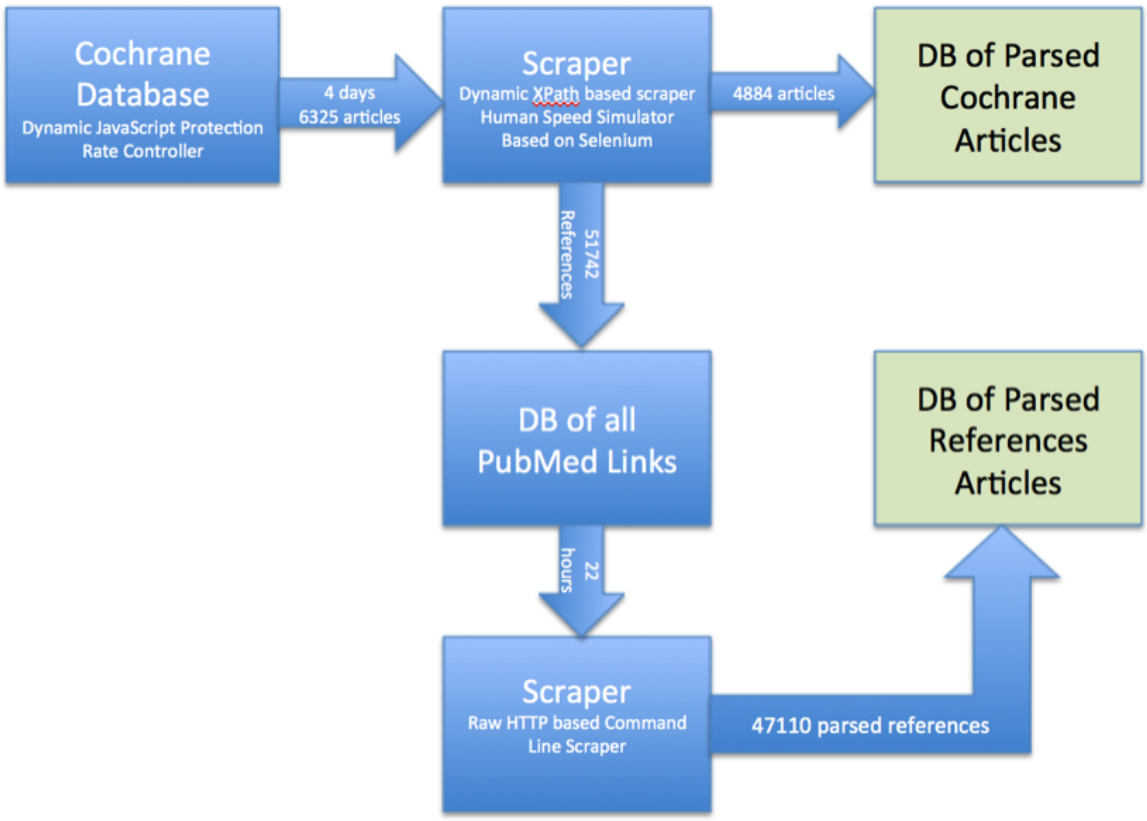

Figure 1: Procedure to Scrape data along with statistics

2.2 Data Filtering

There are three main type of Cochrane Articles, in which are we are only using one of them for this project:

- *Protocols* are brief summary of on-going Cochrane Review, as it was incomplete, protocol articles are omitted from scraping process
- *Comparison-Based* articles compare pros and cons of method X against another method Y
- *Evaluation articles* objective is to determine whether X method or treatment is effective for symptom Y.

Because algorithm is designed to only predict binary YES, and NO decision, we will only consider evaluation articles as a dataset, and excludes all articles of type Comparison-Based, and Protocols. Out of 6,325 articles downloaded, we removed 1,441 articles, which was not evaluation-based reviews.

3 Method

As we saw in previous section, when polarity of Authors conclusion is positive for a Cochrane article, reference articles that has been cited by it also tends to support the claim. Based on this fact, we reasoned that it is possible to predict the final decision as being YES, or NO by analyzing the polarity of referenced articles.

Because of limitation of enough supervised data, it was not possible to train classifier that is able to perform binary classification on polarity of referenced articles. Also, context of referenced articles are rich; therefore, it was hard to capture entire structure with only few hand labeled points.

The number of Cochrane Articles is on the order few thousand while PubMed reference articles are about 50 thousand. Therefore, if we develop a method that is able to use labeled Cochrane Articles to train model, and test the accuracy rather than labeled referenced articles, we need far less number of supervised data. In this project, we will explore two different approaches to achieve this goal:

- *Represent Cochrane Review article* as a combination of reference articles feature vectors
- *Predict reference articles* polarity based on K-Means++ clustering algorithm, and further predict polarity of Cochrane Review article by taking majority vote

3.1 Feature representation of Cochrane Review article

In this method, we will generate feature vectors from reference articles using heuristic based feature extractor algorithm, and represent the parent Cochrane Review article by combining feature vectors from references. We will later use this representation feature vector along with Cochrane Review articles hand labeled tag to train Logistic Regression (LR) model, and test the accuracy.

Heuristic Based Feature Extractor Algorithm converts reference article to feature vector using key phrases, and relative positioning between verbs, and nouns.

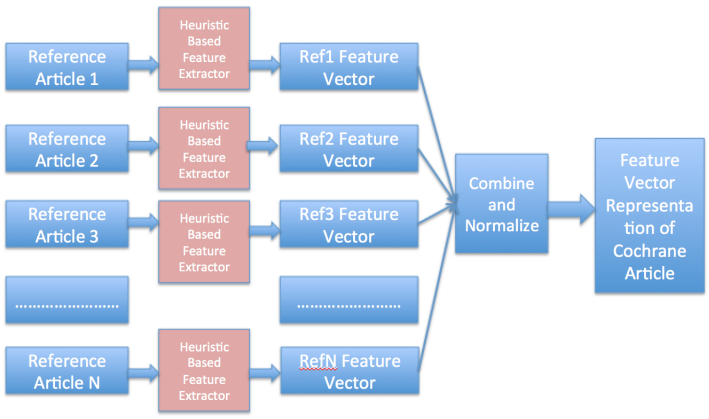

Figure 2: Block Diagram Representation of Algorithm

3.1.1 Heuristic Based Feature Extractor Algorithm

We defined 23 most commonly used phrases, and representing reference article by one-hot feature vector of length 46, where each phrase has two entry on it depending on polarity. If one of the key phrases appears in CONCLUSIONS section of reference article, corresponding entry in feature vector is one depending on polarity of the phrase.

To define the polarity of the individual phrase, we are checking whether there are negation words such as no, and not preceding it within the two words ahead. If there is negation word, we will define polarity of the phrase as being negative.

It is possible that phrase appearing more than one times in the CONCLUSIONS section with different polarity. In that case, we are giving precedence to first appearance, as the style of review is stating most important facts in the first one or two sentences.

Following is the list of key phrases included for extracting the feature vector:

- *Action Verbs*: "support", "improve", "indicate", "effect", "affect", "remove", "reduce", "decrease", "benefit", "less", "increase"
- *Adjectives*: "significant", "potential", "well", "statistical", "strong", "substantial"
- *Nouns*: "difference", "evidence", "better", "unclear", "safe"

Once we convert all reference articles into feature vector form, we added all feature vectors of references articles, and normalized it by dividing total number of reference article cited to represent a Cochrane Review Article. Exact reason for dividing by total number of references rather than normalizing the feature vector will be discussed in Results section.

We also tried bag of words feature vector representation as an alternative to carefully chosen phrase words. We removed most commonly used 15 words, as they are just stop words such as of, in, the, and on. We also excluded words that appeared less than 5 times in the dataset, as there is no significant contribution from these words for polarity. Performance difference between these two approaches will be elaborated in detail in Results section.

3.2 Clustering Based Polarity Detection

It has been noted from dataset that there is similar set of words appear for same kind of polarity. Therefore, it might be possible to cluster references article based on their bag of word feature vector representations discussed in previous section, We used K-Means++ with two clusters and visually inspected random samples from each cluster to determine polarity of articles in cluster.

Even with two clusters each corresponding to positive and negative, and looking at 100 random articles from each, we found that result was extremely accurate, and reliable. (Please refer to clusterResult.p file in github repo to visually inspect the results as well.)

Once we found the polarity of reference articles, we will simply take majority vote from referenced articles and assigns a label for Cochrane Review Article.

Table 1: Example of Clustering Results

Example Articles From Clusters		
Cluster Number	Polarity	Examples
0	NO	To date, adjuvant RT has not been shown to improve overall survival compared with active surveillance.
	NO	Generous financial incentives, as designed in the UK pay for performance policy, may not be sufficient to improve quality of care and outcomes for hypertension and other common chronic conditions.
	NO	The therapeutic efficacy of rTMS was not demonstrated when rTMS was applied to the hand motor cortical area in patients with chronic neuropathic pain at multiple sites in the body, including the lower limbs, trunk, and pelvis.
	YES	Providing information to support choice did not adversely affect attendance for screening for diabetes.
	NO	This regimen of supplementation did not result in consistent improvements in ratings of behavior in lead-exposed children over 6 months.
1	YES	Physiological symptoms and psychological symptoms were both significant predictors of QOL.
	YES	School-based drug prevention programs can prevent occasional and more serious drug use, help low- to high-risk adolescents, and be effective in diverse school environments.
	YES	Neither short-term, medium-term, nor long-term oral creatine supplements induce detrimental effects on the kidney of healthy individuals.
	YES	Our results may encourage the application of extended surgical procedures in patients who would otherwise be rendered incompletely debulked after primary cytoreduction.
	YES	Epo is effective in improving haematological response and reducing RBCT requirements, and appears to have a positive effect on HRQoL.

4 Results

We hand labeled over 320 Cochrane Review Article, and only 129 of them had more than 5 PubMed reference articles, which we considered sufficient to draw conclusion from references.

We trained Linear Regression model using feature vectors, and labels of Cochrane Review Article. To test the accuracy, we performed 10-fold cross validation, and chosen random as a baseline since the task is binary classification.

For clustering method, accuracy is determined based on correctness of majority voting against hand labeled ground truth.

Following is the is the results obtained by experiment:

Table 2: Results of Different Methods

Method Name	Feature Vector	Accuracy	TP	FP	TN	FN
Random	*	0.502	0.253	0.250	0.249	0.247
LR	Defined Phrase	0.598	**0.305**	0.206	0.293	**0.195**
LR with Unit Normalization	Defined Phrase	0.553	0.272	0.227	0.281	0.219
LR	Bag Of Words	**0.607**	0.301	0.194	0.306	0.198
Cluster	Bag Of Words	0.561	0.231	**0.169**	**0.330**	0.269

(a) TP - True Positive, FP - False Positive, TN - True Negative, FN - False Negative

(b) The distribution of the YES and NO label where 49.31%, and 50.69% correspondingly

As we can see from results, there was no significant difference between constructing feature vectors from manually chosen phrases, and bag of words form the dataset. This result suggests that key phrases indeed capture most information about polarity of conclusions.

However, the key difference between two different feature vector representation was how we normalized the sum of reference feature vectors. On the defined-phrase feature vector representation, dividing it by total number of references reached higher accuracy than normalizing vector normally to have length one.

The reason is that the more references that article has, we have to find more number of references that support or refutes the hypothesis. Therefore, with regular normalization, which enforces vector to have unit length, it is taking more into account that which phrase occurred the most rather than occurrence of phrases with relative to number of references. To add relative ordering, we are dividing by total number of references to normalize the vector.

Clustering method achieved the best TN rate, which confirms assertion that clustering algorithm did reasonable classification towards recognizing positive polarized articles. This also indicates that words included in positively polarized references are highly distinctive, and different from rest of the word set.

Overall performance best performance is achieved by LR model with Bag Of Words feature vector. This achieved accuracy is 10% better than our baseline model for randomly guessing binary labels which achieves about 50% of accuracy. False positive rate of best performing LR model is also 6% lower than random input. Therefore, current results prove that it is possible to predict the polarity of summary article for given hypothesis based on polarity of reference articles.

5 Future Work

The main part of this project is conversion of reference articles to feature vector representation. The better feature vector representation is, and the more information we capture with feature vector representation, we expect to see better accuracy with lower false negative and positive rates.

We are planning to improve this feature vector representation by using word embeddings based on deep neural network. The idea is to learn word embeddings that are similar for words with similar polarity. With that representation, we can cluster references more accurately, and can reach higher true positive, an negative rate.

Also, there is noise associated with data. We later discovered that some of the referenced PubMed articles are not considered for Cochrane Review because of mistakes on data aggregation, or methodological errors on experiments. Therefore, we need to filter those references to make accuracy rate much more realistic.