Honours Project – II Report
CDE

Allaparthi Sriteja,
201302139.

Honours report (CDE, Semester : 3-2)

Medical data analysis

Introduction

Main objective of this work is to identify the possible types or a name of the diseases from the general text in some medical blog. For this to achieve :

1) Have to collect the information about all the diseases available in some medical websites.

2) Have to separate the medical named entities from the general text available in the medical blogs or some medical informative websites.

3) Have to process these medical named entities and classify them as signs or symptoms, Disease names, medical treatment procedures, drug names etc.

4) Visualise the symptoms to find out what exactly are common between two diseases. This information is used to classify a disease into a particular class or to gain some knowledge on how to differentiate between two diseases.

5) Train the model to give the most matching disease based on the given list of the symptoms.

Collection of the information about a disease separately

All the information regarding a particular disease is scraped and fed into the system such that this will be used to detect the medical named entities and map them to this disease. This information is functioning like a knowledge engine which will provide the necessary information about a particular disease among the collected ones in the dataset.

Sample collected information about a disease is given here. https://drive.google.com/file/d/0B-8Dn6UcwZpkY1VJSmZfcndObk0/view?usp=sharing

Medical named entity extraction from a given general text

Medical Entity Recognition is a crucial step towards efficient medical texts analysis.

The task of a Medical Name Entity Recognizer is twofold

(i) identification of entity boundaries in the sentences.

(ii) entity categorization.

Our objective is to extend medical entity recognition from the general text.

Medical entities can be diseases, drugs, symptoms, etc. Previously, researchers in the field have used hand crafted features to identify medical entities in medical literature. It has been found that in contrast with semantic approaches which require rich domain knowledge for rule or pattern construction, statistical approaches are more scalable.

Tool used:

Meta-Map

MetaMap is a highly configurable program developed to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text.

Metamap - used because they are domain oriented tags and diseases, symptoms, drugs lie

in corresponding medically specific class clusters - which allows it to identify and categorize

medical terms significantly. Metamap uses a UMLS Metathesaurus dictionary made

specifically for this purpose according to which it assigns tags accordingly.

Definitely useful for NER.

Example :

dsyn for disease, and phsu, imft for drugs, and sosy for symptoms, etc. are significant.

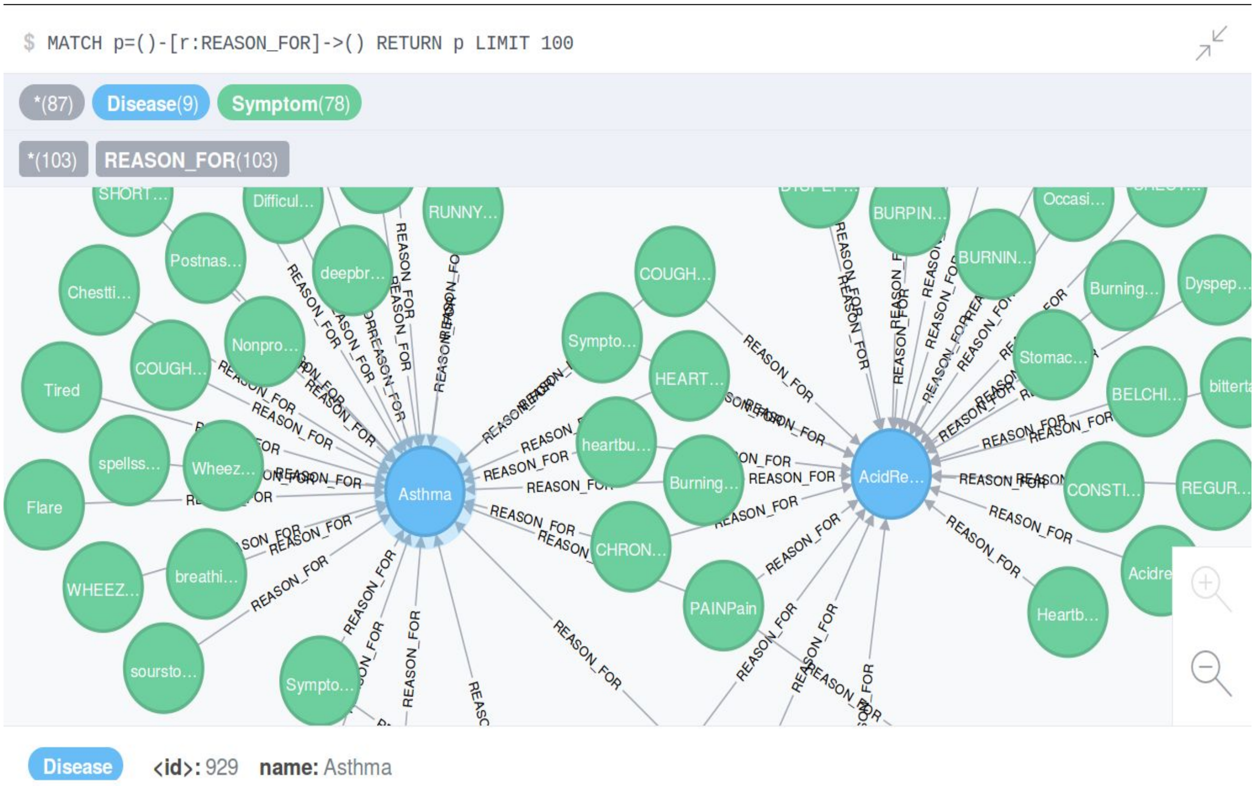
Accuracy of these medical NER : 0.8871

Sample file after Medical Named entity recognition is given here.

https://drive.google.com/file/d/0B-8Dn6UcwZpkN1h4R2x4cnk1SlU/view?usp=sharing

Visualisation of the network of symptoms for all the diseases in the dataset

All the symptoms of a particular disease are collected. Now all these symptoms are mapped for all the diseases to find how many symptoms are common between 2 given

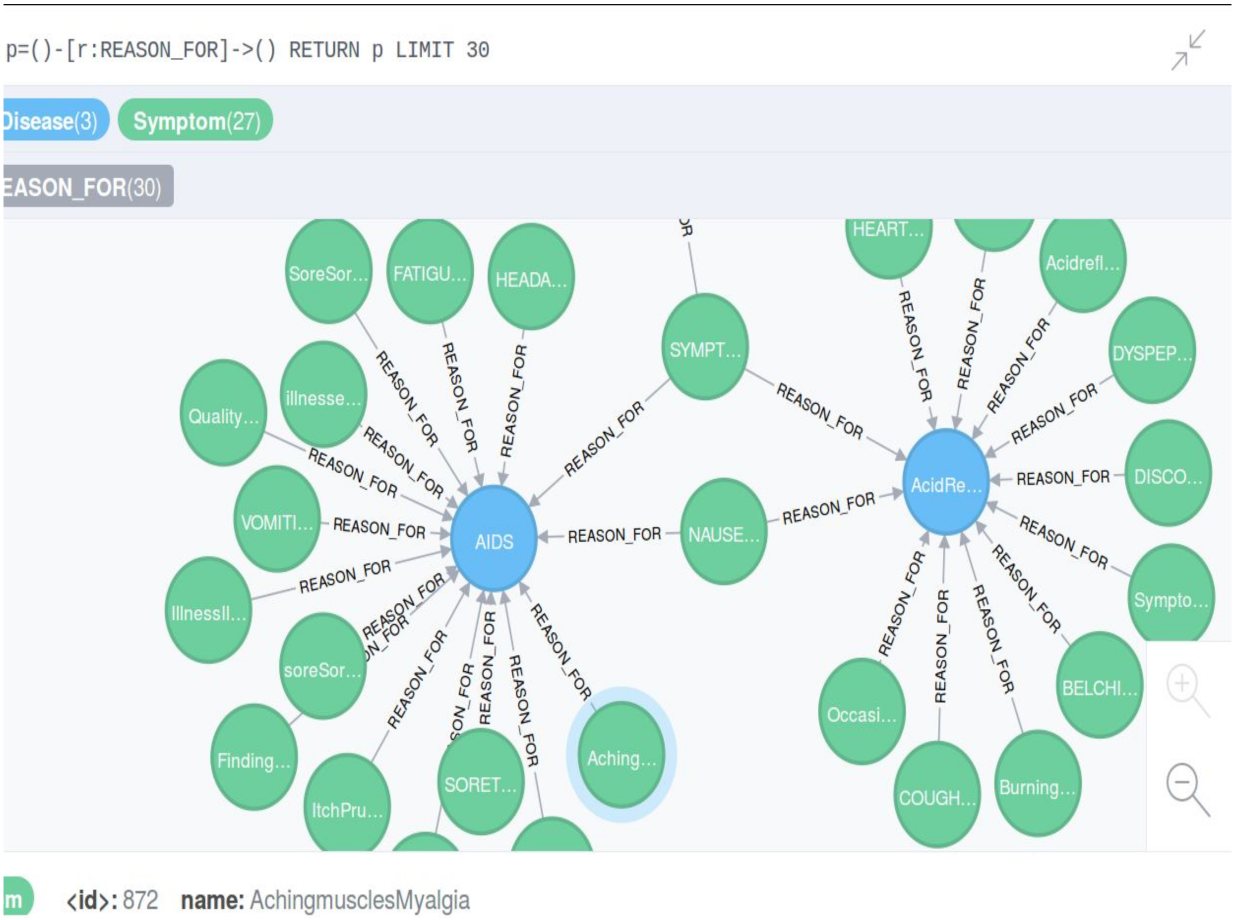


```
$ MATCH p=()-[r:REASON_FOR]->() RETURN p LIMIT 100
```

Disease <id>: 929 name: Asthma

the

diseases or Is there any similarity in the symptoms of the 2 given diseases which will be used to classify the disease into a particular class. An example of the network is shown in the above image for 2 diseases. Here ‘Asthma’ and ‘Acid Reflux’ are the two diseases shown with their symptoms mapped in green colour around the disease. And the common symptoms acts like a bridging nodes between the diseases. You can see the green nodes in middle of the 2 diseases are common for both these two diseases.



```
p=()-[r:REASON_FOR]->() RETURN p LIMIT 30
```

m <id>: 872 name: AchingmusclesMyalgia

Preserving the privacy of medical records and finding similar patients

The main motto of this project is to match the similar records but not by literally matching

the original content. So, here comes the role of preserving the privacy of the original

content of the medical record. Here, we have to generate the hash function such that

similar features should be mapped to similar hash values. From those hash values we

should be able to calculate the similarity of the features.

This project is like an extension to this paper:

http://www.sciencedirect.com/science/article/pii/S1532046415002841

Problem with extending the approach of Bloom Filters (given in PPSM software) for strings to the collection of strings or text :

If the textual data is too large, even though the similar texts don't miss with this approach of bloom filters, but there will be more false positives increasing with the increase in the size of text data after some threshold. So, it will badly affect the precision. Not much useful

in our case.

Brief idea of Working and Implementation :

Consider the huge textual data of a particular record as a single Document. Now, we have the Documents in which the textual data of each record is present separately. In this algorithm, we use a hash function which takes a 32bit integer and maps it to a different integer with almost no collisions.

Hash(x) = (constant 'p' x + constant 'q')%constant 'r'

The constants p and q are randomly chosen such that they should be than the maximum value of 'x'. Constant 'r' is a prime number slightly bigger than the maximum value of 'x'. We can generate different hash functions with different values for the constants p, q. Generate 'k' random hash functions with different values of p,q. Now, apply the first hash function to all the terms of a document and take the minimum hash value generated. Now,

repeat this procedure with all the other k1 hash functions and generate the minimum hash values in each case. These k hash values can be used as a Signature for that particular Document.

Now, get all the k minimum hash values for each document. Then we can get the similarity

of documents by calculating the no. of components in a signature are matching.

Example :

Consider for Document1 Signature1 consists of (‘hash1’, ‘hash2’, ‘hash3’, ‘hash4’)

Consider for Document2 Signature2 consists of (‘hash6’, ‘hash3’, ‘hash1’, ‘hash7’)

There are two values common in between the 2 sets of signatures. So the similarity score can be given as 2/6. This method gives results almost similar results to jaccard index. Like in the above example, the jaccard index is also 2/6.

Future Work :

1) The model that uses the privacy preserving methods of text documents with tracking the similarity of the documents at the same time.
2) Make the model to find the possible diseases based on the given set of symptoms.