Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1221-1231
© Research India Publications
http://www.ripublication.com

Building A Scalable Database Driven Reverse Medical Dictionary Using NLP Techniques

Pragya Tripathi

*Dept. of Computer Science Engineering, Pillai College of Engineering,
New Panvel, University of Mumbai, Maharashtra, India.
Email: 123pragya.tripathi @gmail.com*

Prof. Manjusha Deshmukh

*Dept. of Computer Science Engineering, Pillai College of Engineering,
New Panvel, University Of Mumbai, Maharashtra, India.
Email: mdeshmukh@mes.ac.in*

Abstract

In recent year semantic similarity has a great interest in NLP. With the help of NLP features, we describe the design and implementation of a fully functional system called as reverse medical dictionary in order to achieve the efficiency in fast health treatment consultation system. Reverse medical dictionary allows users to get instant guidance on their health problem through a smart health care system. Simply we say that user can search their diseases from the system through sharing his/her symptoms at any point of time and get instant diagnosis. Here we use many algorithms for achieving our goal.

Keywords: *NLP, WordNet, WSD (word sense disambiguation), Semantic Similarity*

1. INTRODUCTION

Information plays very vital role in modern civilization because firstly we convert data into information and then means of this information we will do amazing invention. Therefore, it is important that the information we are using should be accurate. When we talk about medical diagnosis it is a complicated task that requires operating accurately efficiently and effectively. Medical decision is highly specialized and requires effort and determination to identify diseases which shows similar symptoms or that disease which are rare. Here in medical field misdiagnosis is very harmful. This misdiagnosis may result of many reason it may be inexperience of doctors, ambiguous symptoms and incomplete information. So it is very important that doctor should be experienced and have good knowledge about his/her field.

As we know doctors who are highly experienced classify diseases which are based on differential diagnosis method. Using different diagnosis will help doctors to find root diseases that shows similar symptoms and which is done using knowledge and experience and later they confirmed the diseases by performing various test which give them real picture of disease[4].

2. PROPOSED SOLUTION APPROACH

In the proposed system we are finding various diseases according to user input which is number of symptoms. Here we are using techniques which give us to guess the most appropriate diseases that probably associated with patient's symptoms.

2.1.*System Architecture*

Architecture of reverse medical dictionary is described in figure1 .Where we describe our architecture keeping in mind the design for scalability.

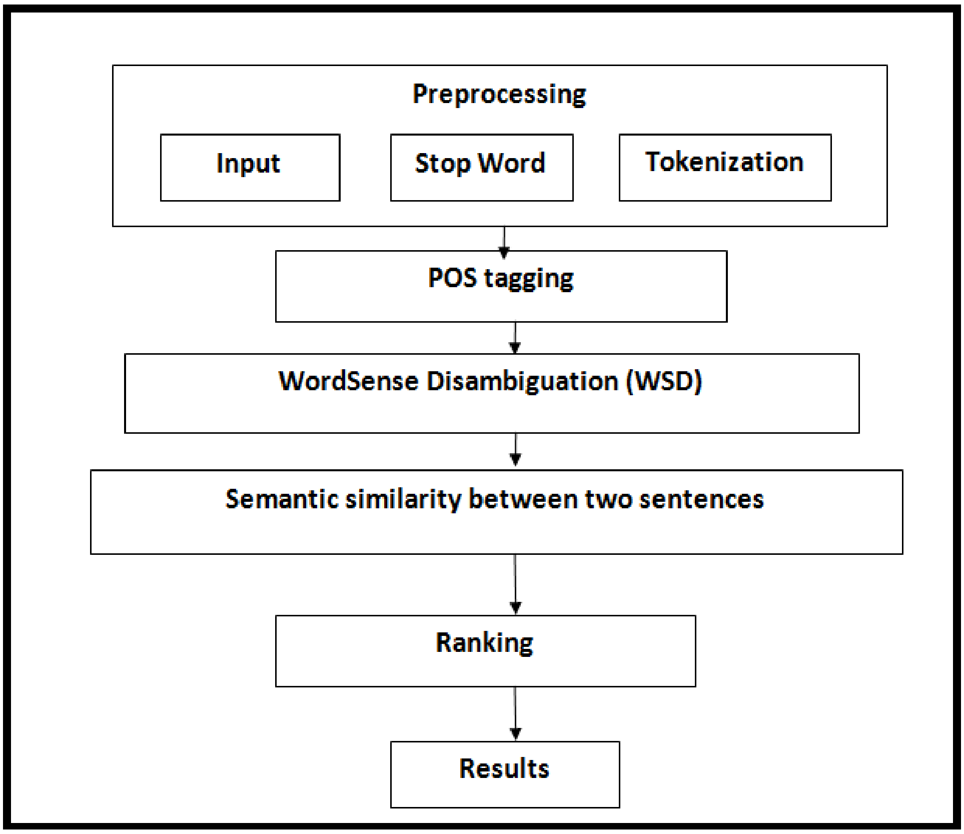

Preprocessing
Input
Stop Word
Tokenization
POS tagging
WordSense Disambiguation (WSD)
Semantic similarity between two sentences
Ranking
Results

Figure 1: Architecture of reverse medical dictionary

2.1.1 Modules for Building Reverse Medical Dictionary

We divide our reverse medical dictionary into number of modules.

1. Preprocessing

a) Stop word removal

Stop words are words which are filtered out after or before processing NLP text. We can say stop words are most common words in any language, although there is no universal list of stop words we can make our own list as per requirement.

b) Tokenization

Tokenization is the process of breaking stream of text or sentence into words, or other meaningful elements called tokens. These tokens become input for further processing.

2. POS tagging

The task of finding correct part of speech (POS - like noun, verb, pronoun, adverb ...) of each word in the sentence is known as tagging. The algorithm takes a sentence as input and a specified tag set (a finite list of POS tags). The output is a single best POS tag for each word. In the project we are plugged in Brill Tagger.

3. WORD SENSE DISAMBIGUATION (WSD)

WSD means the task of finding meaning of an ambiguous word in the given context or the task that identify appropriate meaning(sense) to a given word in a text.

For example Bank can have two senses:

1. Edge of a river 2. Financial institution that accepts money

Here in the project we are using adapted Micheal Lesk algorithm. As the original Lesk algorithm dependent upon glosses (dictionary definition) which are present in traditional dictionary such as oxford. But in adapted form we take advantage of WordNet dictionary which offers highly interconnected set of relation among synonyms. This algorithm is able to compare the glosses of the words that are related to the words to be disambiguous. This will give us richer source of information and it will improve overall disambiguation accuracy. Based on the WordNet hierarchy, the adapted Lesk algorithm takes into account hypernyms, hyponyms, holonymy, meronymy, troponymy, attribute relations, and their associated definitions to build an enlarged context for a given word meaning. Hence, they attempt to enlarge the dictionary-context of a word sense by taking into account definitions of semantically related concepts.

To disambiguate each word in a sentence that has N words, we call the algorithm which is described below.

1. Select a context and for optimization if N is long time, we will define K context around the target word (or k-nearest neighbor) as the sequence of words starting K words to the left of the target word and ending K words to the right.

2. For each word in the selected context, we look up and list all the possible senses of both POS (part of speech) noun and verb.

3. For each sense of a word (WordSense), we list the following relations:

a) Its own gloss/definition that includes example texts that WordNet provides to the glosses.

b) The gloss of the synsets that are connected to it through the hypernym relations.

c) If there is more than one hypernym for a word sense, then the glosses for each hypernym are concatenated into a single gloss string (*).

d) The gloss of the synsets that are connected to it through the hyponym relations (*).

e) The gloss of the synsets that are connected to it through the meronym relations (*).

f) The gloss of the synsets that are connected to it through the troponym relations (*).

(*) All of them are applied with the same rule.

4. Combine all possible gloss pairs that are archived in the previous steps and compute the relatedness by searching for overlap. The overall score is the sum of the scores for each relation pair.

5. Once each combination has been scored, we pick up the sense that has the highest score to be the most appropriate sense for the target word in the selected context space. Hopefully the output not only gives us the most appropriate sense but also the associated part of speech for a word.

*The above algorithm benefit is that it allows user to find the most appropriate sense for each word in a sentence.

4. SEMANTIC SIMILARITY BETWEEN TWO SENTENCES

After doing WSD our last step is finding semantic similarity which is sometimes called as topological similarity. Semantic similarity is calculated at many levels like document level, term level and sentence level. Here we are finding semantic similarity between two sentences, one sentence is input sentence which is symptoms which are going to match with other sentence, means the sentences which are in the databases.

Let us consider the given two sentences as an input to this process; first the words of two sentences are compared. If the two words of the sentences are matched, its similarity score is calculated which are based on syntactic level. If the words of the two sentences are not matched, then synsets of the word is extracted from sentnce1and compared with the other word of the sentence2. If the words are matched at Synset level then return the score as 1, otherwise return 0. Even the words are not matched, then consider the definition of the word sense of the sentences and compare the similarity score of the sentences which are totally based on semantics. This way we compute how two sentences are similar semantically.

Pseudo code for finding semantic similarity between two sentences:

First we build a semantic similarity matrix of each pair of words. If a word does not exist in dictionary we use Edit-Distance Similarity Method.

For making semantic similarity matrix we have to first find out the LCA (least common ancestor).

For finding LCA we have to follow these steps:

```
IF word1 and word2 POS IS NOT EQUAL, then return 0.0

IF word1 and word2 IS EQUAL, then return 1.

ELSE Find, Least Common Ancestor

Assumption: @Distance = 2147483647 (Max value of an Integer)

Iterate through First Synset and compare with other Synset.

IF Found THEN @CA = Value of Synset

@distance1 = Distance of LCA from First Synset

@distance2 = Distance of LCA from Second Synset

@len = @distance1 + @distance2 – 1

If @Distance > @Len

@lcaDepthKey1 = Log10 (@CA)

@lcaDepthRootKey1 = Log10 (root_key_1)

@RootDepthKey1 = @lcaDepthKey1 * 1000 + @lcaDepthRootKey1 = 6976.77

@RootDepth1 = DEPTHMATRIX (@RootDepthKey1)

@lcaDepth1 = @RootDepth1 + 1
```

```
@lcaDepthKey2 = Log10 (@CA)
@lcaDepthRootKey2 = Log10 (root_key_2)
@RootDepthKey2 = @lcaDepthKey2 * 1000 + @lcaDepthRootKey1
@RootDepth1 = DEPTHMATRIX (@RootDepthKey1)
@lcaDepth2 = @RootDepth2 + 1
@LCADepth = @lcaDepth1 + @lcaDepth2
@depthWord1 = @distance1 + @lcadepth1 – 1
@depthWord2 = @distance2 + @lcadepth2 – 1
@Distance = @len
@LCA = CA
Repeat above algorithm until full list of Synsets is iterated.
Calculate R[m , n] using Wu Palmer Method
IF Distance EQUALS 0, then return 1.0
ELSE return (@LCADepth) / (@depthWord1 +@ depthWord2)
```

5. RANKING

From the previous step we got the similarity score .We rank our score in decrease order because we got number of outputs from previous step.

6. RESULTS

After ranking this is the last step in which we have to set one threshold value (here we are taking 0.9).So we take only those result whose score is greater than 0.9 rest we eliminate. This module output is our final results.

Experiment Results :

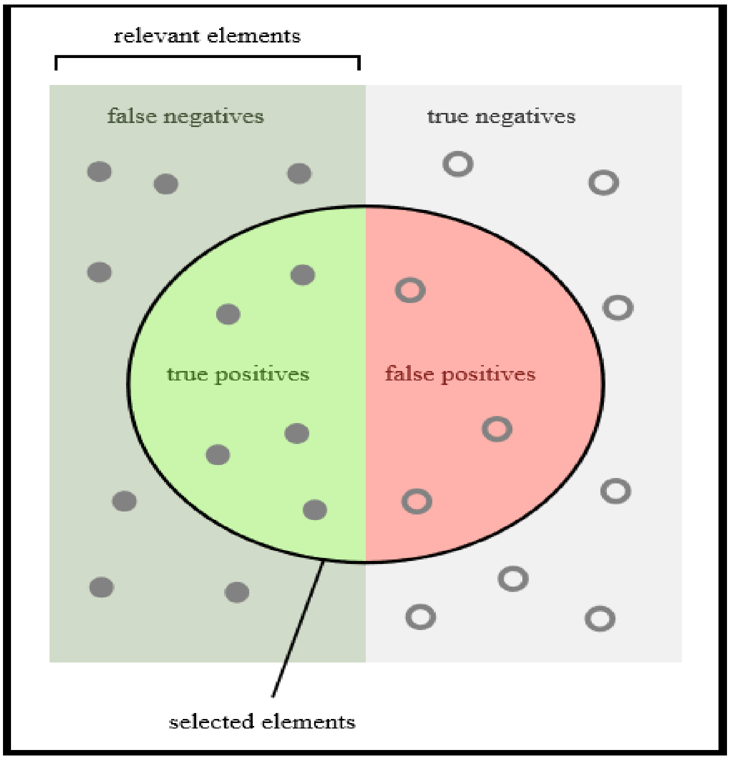
Here we are using databases consisting of 1000 number of diseases. We ran our algorithm on this database. The various similarity score were computed and compared. Given below we are representing the tabular format of test status of different test case.

Table 1. Sample pair of sentences

Inputs	**Obtained Output**	**Test Status**
pain in stomach	Abdominal rigidity, diarrhea, ulcer, q fever.	Identified
redness in eye	Conjunctivitis, corneal abrasion, foreign object in the eye, fungal eye infections.	Identified
bleeding gums	Leukemia, painful gums, talon rodenticide poisoning.	Identified
aching, tiredness, fever	Chicken pox	Identified
extreme pain in legs	-	Unidentified
pain in legs	lymphatic filariasis	Identified
aching, tiredness, fever	Chicken pox	Identified

Result Analysis:

To test our system performance, Actual result analysis of proposed system in terms of precision, recall and accuracy is shown in figure 3.



Precision

Precision is nothing but how many of the returned results are correct, that means the ratio of correct positive observations.

$$\text{Precision} = TP / (TP+FP) \tag{1}$$

Recall

Recall is nothing but how many of the positives the system will return, that means the ratio of correctly predicted positive events. As recall increases precision decreases and vice-versa.

$$\text{Recall} = TP / (TP+FN) \tag{2}$$

Accuracy

Accuracy is perhaps the most intuitive performance measure. It is simply the ratio of correctly predicted observations.

$$\text{Accuracy} = (TP+TN) / (TP+FP+FN+TN) \tag{3}$$

4.1 Result Analysis Graph

Given below is the table which shows the parameters taken to draw a precision, recall and accuracy graph for our proposed system that is reverse medical dictionary.

Table 3: Parameters taken to draw a precision, recall and accuracy graph

Total no of inputs	TP	TN	FP	FN
50	46	2	1	1

Using equation 1,

Precision = TP / (TP+FP) = 46/47 =97%

Using equation 2,

Recall = TP / (TP+FN) =46/47=97%

Using equation 3,

Accuracy = (TP+TN) / (TP+FP+FN+TN)= 46+2/50 =96%

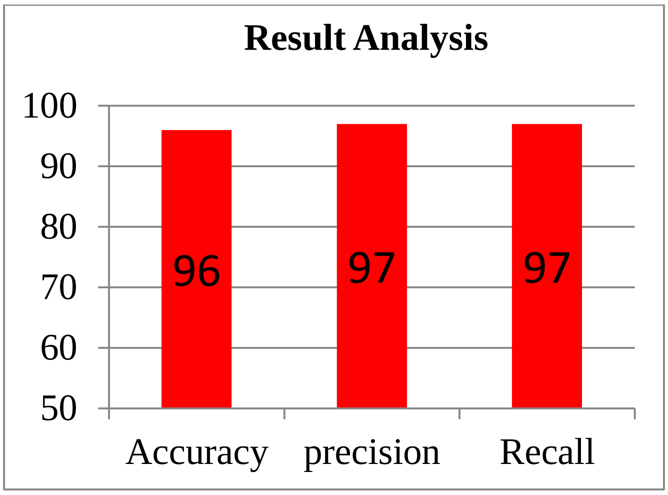


Figure 3: Result analysis of Reverse Medical Dictionary

There are systems available which can perform same work as reverse medical dictionary. By doing rigorous testing with those systems and with proposed system, comparison is shown below in figure 3. Here we are comparing our system by Isabel symptom checker and mayo clinic.

Table 4: System comparison data taken to draw a precision, recall and accuracy graph

Name of the system	Total no of inputs	TP	TN	FP	FN
Reverse Medical Dictionary	20	17	1	1	1
Isabel Symptoms Checker	20	15	2	2	1
Mayo Clinic	20	16	2	1	1

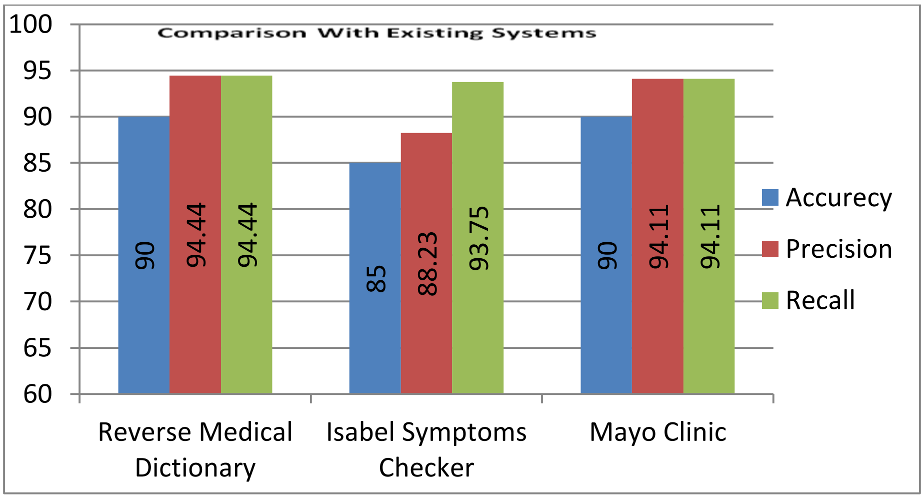


Figure 4: Comparison of accuracy of proposed system with existing systems

7. CONCLUSION

As we know that Reverse medical dictionary has vital role in medical field, the medical dictionary is a system which supports end user and on the other hand it is a consultation project in which users can get the instant guidance on their health problem through an intelligent system which we called as reverse medical dictionary. It will give the user a chance that he/she can change the way to speak your doctor forever. Reverse medical dictionary is the system in which doctors as well as nurses can trust. The system will improve treatment processes and does the detection of disease as earlier as possible. Here the system contains a database that database contains diseases name along with their various symptoms. It also has an option for users or we can say patient to sharing their symptoms. The system processes those symptoms to check for various diseases that can be associated with it. If user's symptoms do not exactly match any disease in the database, then it is shows the diseases user could probably have based on his/her symptoms. We propose a set of methods for building reverse medical dictionary. Here the main concept is finding the similarity measure between the two sentences. As we all know that medical dictionary is available online which uses concept of data mining technique, but we proposed reverse medical dictionary by using NLP concepts which is very different from others medical dictionary. It is a new concept and we describe a set of experiments that show the quality of our results, as well as the runtime performance under load.

8. FUTURE SCOPE

1. Extend the WSD algorithm with supervised learning with such methods as the Naive Bayesian Classifier model.
2. Disambiguate part of speech using probabilistic decision trees.

ACKNOWLEDGMENT

I am thankful to Prof. Manjusha Deshmukh for giving valuable inputs required for the topic, and her invaluable support and also for providing me the idea of this topic .I would like to thanks M.E. Coordinator Prof. Sharvari S. Govilkar and all staff members for their direct and indirect support. I would like to thank the H.O.D. of computer engineering department Dr.Madhumita Chatterjee for her invaluable support.I would like to thank Principal Dr. R. I. K. Moorthy for their invaluable support.

REFERENCES

[1] Sujatha R, "A Survey of Health Care Prediction Using Data Mining" International Journal of Innovative Research in Science, Engineering and Technology Vol. 5, Issue 8, August 2016

[2] Sona Baby, Ariya T.K "A survey paper of data mining in medical diagnosis" International journal of research in engineering and technology, 2014

[3] Ryan Shaw, Anindya Datta, Debra Vander, and Kaushik Dutta, Member, IEEE ," Building a Scalable Database-Driven Reverse Dictionary" IEEE transaction , Vol. 25, no 3, March 2013.

[4] E. Gabrilovich and S. Markovitch, "Wikipedia-Based Semantic Interpretation for Natural Language Processing", Journal of Artificial Intelligence Research, vol. 34, no. 1, pp. 443-498, 2009.

[5] Thabet Slimani, "Description And Evaluation Of Semantic Similarity Measures Approaches", 2007.

[6] T. Pedersen, S. Banerjee, S. Patwardhan: Maximize semantic relatedness to perform word sense disambiguation, 2005.

[7] Thanh Ngoc Dao, Troy Simpson "Measuring Similarity between sentences", 2005

[8] J.P Sutton "Smart medical systems" Nat. Space Biomed. Research. Inst., Houston, TX, USA, 06 January 2003.

Site References

[1] http://dictionary.reference.com/reverse

[2] https://symptomchecker.isabelhealthcare.com/