Relevance Detection and Argumentation Mining in Medical Domain

Vijayasaradhi Indurthi
IIIT Hyderabad, India
vijaya.saradhi@students.iiit.ac.in

Subba Reddy Oota
IIIT Hyderabad, India
oota.subba@students.iiit.ac.in

ABSTRACT

In this paper we describe a method to determine the relevancy of a query with a sentence in the document in the field of medical domain. We also describe a method to determine if the given statement supports the query, opposes the query or is neutral with respect to the query. This is a part of CHIS shared task at FIRE 2016.

Keywords

Information retrieval, argument mining, relevancy detection

1. INTRODUCTION

World Wide Web is increasingly being used by consumers as an aid for health decision making and for self-management of chronic illnesses as evidenced by the fact that one in every 20 searches on google is about health. Information access mechanisms for factual health information retrieval have matured considerably, with search engines providing Fact checked Health Knowledge Graph search results to factual health queries. It is pretty straightforward to get an answer to the query "what are the symptoms of Diabetes" from the search engines. However retrieval of relevant multiple perspectives for complex health search queries which do not have a single definitive answer still remains elusive with most of the general purpose search engines. The presence of multiple perspectives with different grades of supporting evidence (which is dynamically changing over time due to the arrival of new research and practice evidence) makes it all the more challenging for a lay searcher.

2. SHARED TASKS

We use the term "Consumer Health Information Search" (CHIS) to denote such information retrieval search tasks, for which there is "No Single Best Correct Answer"; Instead multiple and diverse perspectives/points of view (which very often are contradictory in nature) are available on the web regarding the queried information. The goal of CHIS track is to research and develop techniques to support users in complex multi-perspective health information queries.

Given a CHIS query, and a document/set of documents associated with that query, the FIRST task is to classify the sentences in the document as relevant to the query or not [4]. The relevant sentences are those from that document, which are useful in providing the answer to the query. The SECOND task is to classify these relevant sentences as supporting the claim made in the query, or opposing the claim made in the query [4].

Example query: Does daily aspirin therapy prevent heart attack?

Sentence 1: "Many medical experts recommend daily aspirin therapy for preventing heart attacks in people of age fifty and above." [Affirmative/Support]

Sentence 2: "While aspirin has some role in preventing blood clots, daily aspirin therapy is not for everyone as a primary heart attack prevention method". [Disagreement/Oppose]

3. DESCRIPTION

For the shared tasks described above, we adopt a deep learning approach for solving them. Deep learning is a method which allows computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined in terms of its relation to simpler concepts. By gathering knowledge from experience, this approach avoids the need for human operators to formally specify all of the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. We use a deep neural network to train the sentences.

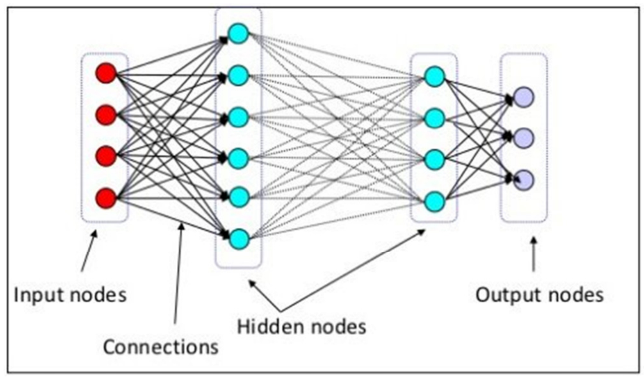

Figure 1. The architecture of a deep neural network

The problems described above are modeled as a supervised learning task [1][4]. For a given query, we have been given a document consisting of a set of sentences. For each sentence we have been provided with the ground truths, i.e. if the sentence is relevant to the query, and if the sentence supports, opposes or is neutral to the query. We have trained a deep neural network [2] for this supervised learning task.

4. FEATURES

We have selected binary bag-of-phrases [3] representation of the document. Since all words in the sentence are not relevant, we have identified the most important features manually and used these phrases to create the feature matrix. Some of the features included the presence of supporting words like 'evidence', 'cause', 'exhibit', 'abnormal', 'nonetheless'. Opposing words like 'oppose', 'does not', 'least', 'less', 'nothing', 'harmless' were also used as features as these words contribute in determining that the sentence opposes the given query. If a feature phrase is present in the given text, the value for that feature would be 1. Otherwise, the value of the feature is 0. All our features are binary. In the preprocessing phase, all text in the upper case was converted to lower case and all numbers were deleted. Some of the feature words and phrases are documented in the table 1.

Table 1. Some relevant phrases used as features

Increase	Intense	Evidence	Harmful
However	Nonetheless	Oppose	Does not
Safe	Healthier	Harmless	Decreased
Inversed	Weak	Deadly	Cancer
Disease	Overdose	Dangerous	Risk
Adverse	Hazard	Poison	Prohibit
Overdose	Irritate	How safe	Associated
Suppress	Side effect	Oppose	Disorder
Incidence	Deficit	Though	Whereas
Nonetheless	Shorten	Reduce	Prevent
Protect	Wards off	Effective	Fewer
Questionable	Benefit	Disagree	Unsupported
Not recommend	Inconclusive	Unjustified	Myth
Viral	Evidence	No increase	Good choice
Flawed	Counteract	Lessen	Cause pain
Still high	Effective	Bothersome	No longer
Inadvisable	Strengthens	Lessens	Fighting
Unlikely	Still high	Good choice	Alarming

Table 2 shows the number of features used for each dataset

Table 2. Features used for each dataset

Query	Number of Features
Q1- Skincare	81
Q2-MMR	64
Q3-HRT	105
Q4-ECIG	95
Q5-Vit C	124

5. ARCHITECTURE

We use a deep neural network for training for both the tasks. The input layer had as many neurons as the input features. Task 1 is a binary classification problem, indicating if the sentence was relevant to the query or not. Task 2 is a multi-class classification problem, which indicates if the sentence supports, opposes or is neutral to the query. Table 3 shows the architecture of the neural network for both of the CHIS tasks [2][5].

Table 3. Neural Architectures for CHIS tasks 1 and 2

Task	Hidden Layers	#Neurons in Hidden layer	Activations
Task 1	2	120, 8	relu, sigmoid
Task 2	2	150, 150	tanh, tanh

For task 1, the classification is a binary classification problem with a binary cross entropy layer at the output. For task 2, it is a multi-class classification problem, and hence a softmax layer is used at the output layer. For training the deep neural network, we used keras. Keras is an open source neural network library written in Python. It is capable of running on top of either Tensorflow or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being minimal, modular and extensible.

We train both the neural networks for 150 epochs for convergence.

6. RESULTS

The following are the results obtained on the test set. Table 4 shows the average precision, recall and F1 score of the classifier for task 1**.** Table 5 shows the average precision, recall and F1 score of the classifier for task 2**.**

Table 4. Task 1 precision on test set

Task	Precision	Recall	F1-score
Q1- Skincare	0.80	0.78	0.78
Q2-MMR	0.84	0.79	0.81
Q3-HRT	0.89	0.89	0.89
Q4-ECIG	0.79	0.66	0.68
Q5-Vit C	0.73	0.73	0.71

Table 5. Task 2 precision on test set

Task	Precision	Recall	F1-score
Q1- Skincare	0.76	0.74	0.75
Q2-MMR	0.55	0.45	0.47
Q3-HRT	0.66	0.54	0.53
Q4-ECIG	0.54	0.52	0.52
Q5-Vit C	0.52	0.50	0.49

7. OBSERVATIONS

Predicting the relevance and determining if a sentence supports the given query is not a trivial problem and needs knowledge of Natural Language Processing and Information Retrieval techniques. In this paper we proposed a fast deep learning method to predict the same using a deep neural network. We observe that the average precision for task 1 is 77.03% and for task 2 is 54.86%. Task 2 is a multi-class problem and is more difficult than task1.

8. FUTURE WORK

In this paper, we have used a select set of phrases as features. Since the sentences and the query, both are short text segments, features using Natural Langauge Processing like POS tagging etc can be used as features augmented with the existing features to improve the precision and recall [6]. Although we have identified the features manually, the features could have been figured out by selecting the adjectives and adverbs using any of the existing NLP toolkits. This would make the solution scalable and generic and can be applied for other similar datasets.

9. CODE

All the code is available at https://github.com/saradhix/chis for research and academic purpose.

10. REFERENCES

[1] Robert Gaizauskas, Mark Hepple, and Mark Greenwood. 2004. Information retrieval for question answering a SIGIR 2004 workshop. *SIGIR Forum* 38, 2 (December 2004), 41-44. DOI=http://dx.doi.org/10.1145/1041394.1041403.

[2] Yoshua Bengio. 2009. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.* 2, 1 (January 2009), 1-127. DOI=http://dx.doi.org/10.1561/2200000006.

[3] Maria Fernanda Caropreso and Stan Matwin. 2006. Beyond the bag of words: a text representation for sentence selection. In *Proceedings of the 19th international conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence* (AI'06), Luc Lamontagne and Mario Marchand (Eds.). Springer-Verlag, Berlin, Heidelberg, 324-335. DOI=http://dx.doi.org/10.1007/11766247_28.

[4] Andrenucci, A., 2008. Automated Question-Answering Techniques and the Medical Domain. In *HEALTHINF (2)* (pp. 207-212).

[5] Andreas Buja and Werner Stuetzle and Yi Shen. 2005. *Loss Functions for Binary Cross Probability Estimation and Classification: Structure and Applications.* Doctoral Thesis. University of Pennsylvania.

[6] Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the OpQA corpus. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT '05). Association for Computational Linguistics, Stroudsburg, PA, USA, 923-930. DOI=http://dx.doi.org/10.3115/1220575.1220691.