Consumer Health Information System

Raksha Sanjay Jalan
Search and Information Extraction Lab
IIIT Hyderabad
Hyderabad, India
jalan.raksha@research.iiit.ac.in

Pattisapu Nikhil Priyatam
Search and Information Extraction Lab
IIIT Hyderabad
Hyderabad, India
nikhil.pattisapu@research.iiit.ac.in

Vasudeva Varma
Search and Information Extraction Lab
IIIT Hyderabad
Hyderabad, India
vv@iiit.ac.in

ABSTRACT

World Wide Web acts as one of the major sources of information for health related questions. However, often, there are multiple conflicting answers to a single question and it is hard to come up with "a single best correct answer". Therefore, it is highly desirable to identify conflicting perspectives about a particular question (or topic). In this paper, we have described our participation in Consumer Health Information System(CHIS) task at FIRE 2016. There were two sub-tasks in this contest. The first sub-task deals with identifying if a particular answer is relevant to a given question. The second sub-task deals with detecting if a particular answer agrees or refuses the claim posed in a given question. We pose both these tasks as supervised pair classification tasks. We report our results for various document representations and classification algorithms.

Keywords

Pair classification tasks, document representations

1. INTRODUCTION

Most of the research developments in area of Question Answering(QA), as fostered by TREC, have so far focused on open-domain QA systems. Recently however, the field has witnessed a growing interest in restricted domain QA.

The health domain is one of the most information critical domains in need of intelligent Question Answering systems that can effectively aid medical researchers and health care professionals in their daily information search.

The proposed CHIS task investigates complex health information search in scenarios where users search for health information with more than just a single correct answer, and look for multiple perspectives from diverse sources both from medical research and from real world patient narratives.

Given a CHIS query,a document/set of documents associated with that query, the task is to classify the sentences in the document as relevant to the query or not. The relevant sentences are those from that document, which are useful in providing the answer to the query. These relevant sentences need to be further classified as supporting the claim made in the query, or opposing the claim made in the query.

We pose both these problems as pair classification tasks, where given a (question, answer) pair, the system has to judge whether or not the answer is relevant to the query and if so, whether or not it supports the claim made in the query. Consider the following example

Question: Are e-cigarettes safer than normal cigarettes?

Sentence 1: Because some research has suggested that the levels of most toxicants in vapor are lower than the levels in smoke, e-cigarettes have been deemed to be safer than regular cigarettes.

Sentence 2: David Peyton, a chemistry professor at Portland State University who helped conduct the research, says that the type of formaldehyde generated by e-cigarettes could increase the likelihood it would get deposited in the lung, leading to lung cancer.

Sentence 3: Harvey Simon, MD, Harvard Health Editor, expressed concern that the nicotine amounts in e-cigarettes can vary significantly.

In the above example Sentence 1 is Relevant and supports the claim made in the question. Sentence 2 is relevant but refutes the claim made in the question. Sentence 3 is irrelevant to the question. For both the tasks, we used K-fold cross validation technique to evaluate our results.

2. RELATED WORK

Our proposed method solves question answering task as classification task.Lot of research work has been done on text categorization.

Text representation is one of the key factors that affects the performance of classifier. The Paragraph Vector algorithm by Le and Mikolov[5]also termed paragraph2vec is a powerful method to find suitable vector representations for sentences, paragraphs and documents of variable length. The algorithm tries to find embeddings for separate words and paragraphs at the same time through a procedure similar to word2vec. De Boom, Cedric and Van Canneyt[1] were first to come up with hybrid method for short text representations that combines the strength of dense distributed representations with the strength of tf-idf based methods to automatically reduce the impact of less informative terms.According to this paper, combination of word embeddings and tf-idf information leads to a better model for semantic content within short text fragments.

Ruiz, Miguel E and Srinivasan, Padmini[8] presented the design and evaluation of a text categorization method based on the Hierarchical Mixture of Experts model. This model

has used a divide and conquer principle to define smaller categorization problems based on a predefined hierarchical structure. The final classifier was a hierarchical array of neural networks. They have shown that the use of the hierarchical structure improves text categorization performance with respect to an equivalent flat model.
Dumais, Susan[2]has experimented with different automatic learning algorithms for text classification.Each document is represented as vector of words as done in vector representation of information retrieval[9].This vectros are then fed to different classifiers for text categorization.Experiments have shown that Linear Support Vector Machines(SVM) is more promising as compared to other classifiers on their dataset.But for our task Naive Bayes has outperformed.

3. APPROACH

In the pair classification task, i.e. categorizing the pair (q_m, a_n) we create two labeled datasets for each query as shown below.

$$RelevanceDataset_{q_m} = \{(a_n, 1)\ such\ that\ a_n\ is\ relevant\ to\ q_m\} \cup \{(a_n, 0)\ such\ that\ a_n\ is\ not\ relevant\ to\ q_m\} \quad (1)$$

$$ClaimDataset_{q_m} = \{(a_n, 1)\ such\ that\ a_n\ supports\ the\ claim\ made\ in\ q_m\} \cup \{(a_n, 0)\ such\ that\ a_n\ refutes\ the\ claim\ made\ in\ q_m\} \cup \{(a_n, 2)\ such\ that\ a_n\ is\ neutral\ to\ the\ claim\ made\ in\ q_m\} \quad (2)$$

Note that we could use the above dataset creation techniques only because the number of questions were fixed and known in advance.

We observed that, labels were highly imbalanced in both datasets with a larger number of positive examples and fewer negative examples. We use oversampling and under sampling based techniques to mitigate this problem (OverSampling technique:Synthetic Minority Over-sampling Technique (SMOTE)). After creating the datasets. We split the data into train and test sets. We use doc2vec and tf-idf and ensemble based representations to represent each answer (or sentence). We train multiple supervised algorithms on each of the above mentioned datasets.

3.1 TF-IDF

TF-IDF representation is one of the well established document representation technique in the field of text mining. This kind of representation is capturing syntactic similarities as for the example *(is cancer curable?, Chemotherapy is often used to cure cancer)*. However, TF-IDF based representations are not efficient at capturing the semantic similarities between sentences as in the example: *Does sun exposure cause skin cancer ?, Exposure to UV rays from the sun or tanning beds is the most preventable risk factor for melanoma*. Note that *melanoma, cancer* are highly similar concepts but their similarity is not captured in TF-IDF representation. We therefore also experiment with representations that are good at capturing the semantic relations between text. We have used the TF-IDF implementation of scikit-learn.

3.2 Doc2Vec

Recently, Word2Vec[6] based models have been exploited heavily for several tasks that require capturing semantic relatedness between text. Doc2Vec[5] is one such model which is trained on huge text corpora for the task of word prediction. The doc2vec algorithm has two variants - Distributed Memory (DM) and Distributed Bag of Words (DBoW). For this work, we use Distributed Memory (DM) based models due to its superior performance in previously reported tasks. The architecture of DM is shown in figure 1

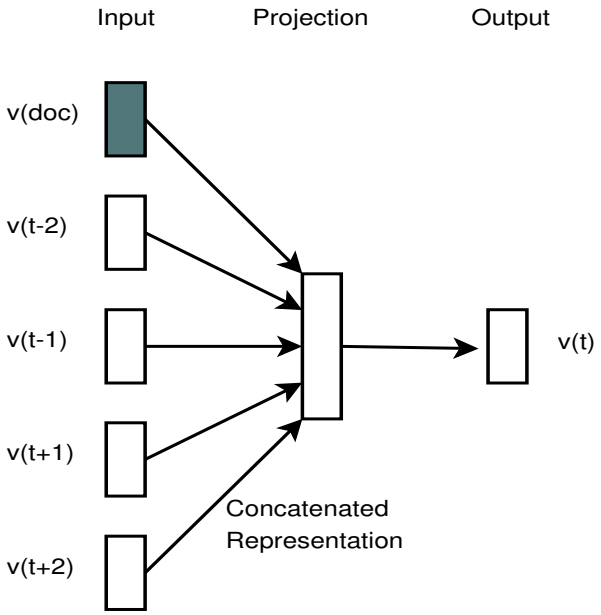

Figure 1: Architecture of Distributed Memory(DM) Model

The problem with doc2vec or any other neural network based model is that it requires huge amount of training data. The main reason for this is the large number of parameters which need to be learnt. Consider the example of doc2vec model shown in figure 1. The vector representations of 4 words, document representation, neural network weights, all have to be learnt. The number of sentences available in CHIS task is too low for such representation learning schemes. To address this issue, we choose pre-trained word vectors which already capture semantic relatedness between words to a large extent.

Although, google released word vectors trained on google news corpus using the word2vec algorithm, we did not choose these vectors as the number of hits were too low. The main reason for this is the difference in domain (many words in the health care domain, found in the CHIS dataset were not present in the google news dataset). We therefore used the vectors released by Pyssalo et al who also train word2vec algorithm on PubMed corpus. We used Gensims implementation for Doc2Vec[1].

3.3 Ensemble Representation

[1]https://radimrehurek.com/gensim/models/doc2vec.html

In order to capture both the syntactic and semantic similarities efficiently, we use an ensemble approach, where for each sentence we obtain its TF-IDF and doc2vec representations (from previous sections). We then concatenate both these representations to form an ensemble representation.

4. DATASET

This CHIS dataset consists of 5 health related queries and 5 files containing labeled sentences for respective queries. Each sentence has two associated labels

- Relevance Label (Relevant or Irrelevant)
- Support Variable (Support, Oppose or Neutral)

The queries are of the following formats, where A, B represent medical entities.

- Does A causes B?
- Does A cure B?
- Is A is better than B?

5. EXPERIMENTS

We used document embedding size of 400 for all the experiments involving doc2vec, word embedding size obtained using word2vec was 200. We have used Pythons sklearn library to realize the SVM, Naive Bayes algorithms.We have realized a neural network using Keras library [2] using Theano as backend. We have used sigmoid as activation function and Binary Cross Entropy(BCE) as loss function. Data is fed to the network in mini-batches with a mini-batch size of 32. We use a 10 fold cross validation to evaluate all our results.

6. RESULTS

In this section we present the results of various document representations and classification algorithms for both the CHIS subtasks: predicting relevant answers and predicting whether or not a given answer supports the claim made in the question.

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	14.62	28.65	48.72
MMR	8.45	21.841	61.762
HRT	10.11	30.54	47.67
E-cigarettes	17.79	21.67	41.985
Vitamin C	6.05	23.45	41.567
Average Accuracy	11.404	25.2302	48.3408

Table 1: Results obtained for sub-task 1 for Doc2Vec representations

For Both the sub-tasks, highest average accuracies are achieved when sentences are represented using ensemble representations and classifications are done using Naive Bayes classifier.

[2]https://keras.io/keras-deep-learning-library-for-theano-and-tensorflow

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	9.76	46.67	57.65
MMR	7.42	30.34	74.862
HRT	9.192	25.43	62.05
E-cigarettes	12.41	25.21	54.785
Vitamin C	7.05	32.51	54.28
Average Accuracy	9.166	32.032	60.725

Table 2: Results obtained for sub-task 1 for TF-IDF representations

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	28.66	62.91	68.181
MMR	12.35	36.06	87.931
HRT	15.92	34.32	75
E-cigarettes	20.81	52.23	71.875
Vitamin C	19.76	50.67	62.162
Average Accuracy	19.5	47.238	73.030

Table 3: Results obtained for sub-task 1 for Ensemble representations

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	26.45	54.95	57.74
MMR	17.67	25.42	49.851
HRT	14.95	24.67	21.56
E-cigarettes	16.67	32.96	41.65
Vitamin C	11.96	35.78	31.41
Average Accuracy	17.54	34.756	40.442

Table 4: Results obtained for sub-task 2 for Doc2Vec representations

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	28.96	57.65	59.54
MMR	19.45	25.24	62.89
HRT	18.65	29.56	35.42
E-cigarettes	17.45	39.567	55.645
Vitamin C	21.05	47.671	31.94
Average Accuracy	21.112	39.817	49.087

Table 5: Results obtained for sub-task 2 for TF-IDF representations

Query Name	Neural Network	SVM	Naive Bayes
Skin Cancer	34.79	60.67	62.5
MMR	21.676	29.508	68.96551724
HRT	21.25	34.66	37.5
E-cigarettes	19.345	46.26	60.9375
Vitamin C	22.197	50.66	32.43243243
Average Accuracy	23.851	44.35	52.467

Table 6: Results obtained for sub-task 2 for Ensemble representations

7. CONCLUSION AND FUTURE WORK

In this work, we have designed algorithms to detect if an answer is relevant to a particular health query and whether or not it supports the claim made in the query. We pose both

these tasks as classification tasks. We experimented with a combination of several document representation schemes and classification algorithms. We note that Naive Bayes classifier has outperformed other classification algorithms by a significant margin. We got the average accuracy of 73.03% in sub-task 1 and 52.46 in sub-task 2. We also additionally note that our model has predicted results with highest accuracy for MMR query. The choice of training one classifier for a query also gave superior performance compared to training one classifier per class. We observed that our model's performance is highly sensitive towards towards quality of pre-trained word vectors, choice of classifier.

We wish to further extend this work by obtaining pre-trained word vectors using other neural network based algorithms like GLoVE[7], Skip thought[4], Deep Structured Semantic Model(DSSM)[3], Convolutional Deep Structured SemanticModels(CDSSM)[10]. We also wish to use these algorithms in order to obtain richer document representations. In this work, we have trained one classifier per query, but such a setting is not feasable for building real applications where the queries are not known in advance. In such scenarios we wish to categorize queries and train a single classifier for each query category.

8. REFERENCES

[1] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt. Learning semantic similarity for very short texts. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234. IEEE, 2015.

[2] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM, 1998.

[3] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM, 2013.

[4] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[8] M. E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.

[9] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[10] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM, 2014.