AMRITA_CEN@FIRE 2016: Consumer Health Information Search using Keyword and Word Embedding Features

Veena P V, Remmiya Devi G, Anand Kumar M, Soman K P Center for Computational Engineering and Networking (CEN) Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham, Amrita University, India

ABSTRACT

This work is submitted to Consumer Health Information Search (CHIS) Shared Task in Forum for Information Retrieval Evaluation (FIRE) 2016. Information retrieval from any part of web should include informative content relevant to the search of web user. Hence the major task is to retrieve only relevant documents according to the users query. The given task includes further refinement of the classification process into three categories of relevance such as support, oppose and neutral. Any user reading an article from web must know whether the content of that article supports or opposes title of the article. This seems to be a big challenge to the system. Our proposed system is developed based on the combination of Keyword based features and Word embedding based features. Classification of sentences is done by machine learning based classifier, Support Vector Machine (SVM).

CCS Concepts

•Information systems \rightarrow Clustering and classification; •Applied computing \rightarrow Health care information systems;

Keywords

Word embedding, Machine Learning, Support Vector Machine (SVM)

1. INTRODUCTION

Natural Language Processing plays a vital role in the interpretation of human language in the most understandable format to the system. This type of role finds application in delivering the most relevant information through web search. Nowadays, information regarding health issues is one among the essential need for people. The number of researches and evidences in this field are growing rapidly day by day. So, if an individual searches through web for any health related query, a large number of documents will be retrieved. The efficiency of the search lies in the fact that the retrieved documents are relevant to the query. The main objective of proposed system is that a person irrespective of his absence of domain knowledge is supposed to get benefited through web search.

In the past years, many developments were made on efficient retrieval of relevant clinical data. A paper was published which discusses on the role of shared task on overcoming barriers to NLP in clinical domain [4]. Medical Records Track was a task conducted for comparing algorithms used for text retrieval for clinical studies by the Text Retrieval Conference (TREC) in 2011 [12]. The ShARe/CLEF eHealth evaluation lab conducted a task to analyze the effect of using additional information like discharge summaries and external resources such as medical ontologies on the Information Retrieval [5]. PubMed, an archive of biomedical journal and Google Scholar were compared to analyze the retrieval efficiency for clinical searches which is discussed in [10]. Embedding features also can be used efficiently for entity extraction which also helps in improving retrieval of precise documents. Several methods were proposed in various FIRE tasks to perform entity extraction such as Conditional Random Field (CRF) based entity extraction [9], Entity extraction for Indian languages using SVM based classifier [2] and Named Entity Recognition for Indian Languages using rich features [1]. A paper was also published on extracting entities for Malayalam language using Structured skip-gram based embedding [8].

The proposed system is useful in finding sentences relevant to the given query and also to find whether it is a supporting, opposing or neutral sentence. An example of training data of the query 'Skin Cancer' is given in Table 1.

The given dataset along with some additionally collected clinical documents from Web were subjected to unsupervised feature extraction. Different approaches like Keyword based and Word embedding based information search were carried out. The integration of these two features achieved better results. Our proposed system is developed based on the combination of keyword and word embedding features. The embedding vectors with the keyword feature vector and the corresponding labels of Relevant/Irrelevant and Support/Oppose/Neutral tags were together given to train classifier. A well-known machine-learning based classifier, Support Vector Machine (SVM) is used for classification task. Section 2 includes the task description. The details about the dataset used is given in Section 3. Section 4 discusses about our proposed methodology. Experimentation and Results are explained in Section 5. The conclusion of the paper is given in Section 6.

2. TASK DESCRIPTION

Our system is submitted in Consumer Health Information Search in FIRE2016 [11]. The task given includes two subtasks. The first task is to classify the sentences in the document as relevant to the query or not (R/IR). The second task is to further classify Relevant and Irrelevant sentences into Support, Oppose and Neutral (S/O/N) sentences with respect to the query. Dataset contains 5 queries say, Q_{skin} ,

Table 1: Example from training data

Sentences	R/IR	S/O/N
Most skin cancers are caused by exposure to the sun.	R	S
Skin cancer can look many different ways.	IR	Ν
Evidence shows that the Sun Protects You from Melanoma.	R	0

Table 2: Number of sentences in Train Data and Test Data

	Train Data								
Query		Relev	ant			Irrelev	vant		Test Data
	Support	Oppose	Neutral	Total	Support	Oppose	Neutral	Total	
Skin Cancer	104	76	13	193	0	2	146	148	88
E-cigarette	93	165	35	293	0	0	120	120	64
MMR-Vaccine	72	92	44	208	0	0	51	51	58
Vitamin- C	111	68	29	208	0	0	70	70	74
Women-HRT	41	132	31	204	1	4	37	42	72

Table 3: Additional Dataset Collected

Query	Additional dataset
Skin Cancer	1044
E-cigarette	1003
MMR-Vaccine	1084
Vitamin-C	1199
Women-HRT	1469

 $Q_{ecig}, Q_{mmr}, Q_{vitc}$ and Q_{hrt} . Each query contains 200 to 400 sentences. For a query say Q_{skin} , each sentence in that particular document is classified for relevant/irrelevant and support/oppose/neutral (given two task) with labels L_1 and L_2 such that $L_1 \in \{R, IR\}$ and $L_2 \in \{S, ON\}$. Thus, each sentence q_i in the document of Q will have two labels - l_1 denoting relevant or irrelevant and l_2 denoting support, oppose or neutral. Relevant sentences are useful in providing answer for the given query. With the help of resulting predicted label from task 1, classification in task 2 has been carried out.

3. DATASET DESCRIPTION

Training data given for this task holds 5 queries Q_1 , Q_2 , Q_3 , Q_4 and Q_5 corresponding to Skin Cancer, E-cigarette, MMR-Vaccine, Vitamin-C and Women-HRT. Each query contains sentences under two categories - Relevant and Irrelevant tags (R / IR tag). For each query, the number of relevant and irrelevant sentences is different. It is further categorized into Support, Oppose and Neutral (S/O/N tag). Individual count of R/IR sentences and S/O/N sentences, for each query is tabulated in Table 2. Additional dataset related to these 5 queries were collected from online resources. Individual sentence count of additional dataset for each query is given in Table 3. Due to time-constraint, we limited the collection of additional dataset around 1000.

After analysis of given training data it has been found that, if a sentence is irrelevant to a query then most probably it will be a neutral sentence with respect to that query. At the time of generation of embedding features, training data and additional dataset is used to train the word embedding model.

4. METHODOLOGY

Our proposed system is based on the combination of Word embedding and Keyword based generation of features. Word embedding features are generally word vectors obtained using the word2vec tool [7]. In this case, the input is sentences of each query. So we need to get embedding features for sentences. These embedding features are obtained from word2vec features. Input for word2vec includes training data and additional dataset collected from online resource. Word2vec is trained to get embedding features for training data. The size of vector is set to 100. Skip gram model is chosen to train word2vec. The embedding features resulting from word2vec is used to generate embedding feature for sentence as in Eq. (1) [6].

$$y = a + Th(D, w_{t-k}, ..., w_{t+k}; W)$$
(1)

where a,T are the softmax parameters and h is the combination of word and sentence embedding features. W stands for word vectors and D stands for sentence vectors. Hence these embedding features are considered to be a feature set for the approach using word embedding model.

The second feature set in our methodology is keyword features. Keywords are extracted from the dataset given for training and testing. For the task to classify the sentences as relevant or irrelevant, keywords are extracted based on its frequency of occurrence in the relevant and irrelevant sentences. The threshold value for determining frequency of words is set as 7 for task 1 and 6 for task 2.

The list of keywords extracted for task 1 and task 2 is given in Table 4. For n keywords, a vector of length n is defined which indicates the presence or absence of the keyword as 1 or 0 respectively. The vector of length n is considered to be the keyword feature in our system. Word embedding model and keyword based model were also separately evaluated using SVM classifier.

Table 4: List of keywords extracted for Task 1 and Task 2

QUERY	KEYWORDS		
QULIU	Task 1	Task 2	
Vitamin -C	Vitamin, Prevent, Symptoms, Severe, Incidence, Dose, Cold	Prevent, Reduce, Severe, Benefit, Risk, No	
E-Cigarette	Smokers, E-Cigarette, Cigarette, Tobacco, Cancer, Quit, Cessation	Safe, Less, Harm, Damage, Risk, No	
MMR-Vaccine	Vaccine, MMR, Autism, Children, Disorder, Thimerosal, Measles	No, Evidence, Cause, Possible, Risk, Develop	
Skin Cancer	UV, Melanoma, Exposure, Cancer, Sun, Skin, Radiation	Increase, Cause, Work, Rate, Exposure, Not	
Women-HRT	Menopause, HRT, Hormone, Ovarian, Breast, Estrogen, Oestrogen	Increase, Effect, Severe, High, Risk, Symptom	

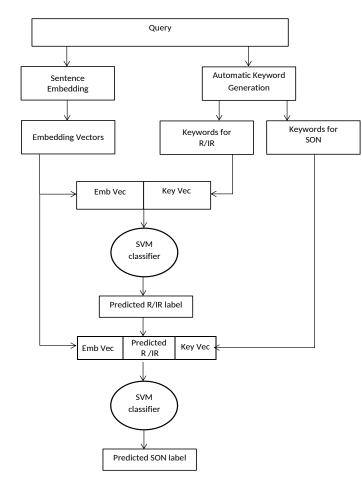


Figure 1: Methodology of the Proposed System

5. EXPERIMENTS AND RESULTS

As mentioned above, our system is developed based on the combination of word embedding and keyword features. The methodology of the proposed system is illustrated in Figure 1. The sentence vectors and keyword features of each sentence in a particular query in the training data are combined. The Relevant/Irrelevant label from the training data is taken. Machine learning based SVM classifier is used for training the system [3]. The combined feature set and the label set is given as input. During training, each query holds a model that includes word embedding and keyword features. Hence there will be 5 models (to classify the sentences into relevant or irrelevant) for 5 queries. These 5 models for 5 queries are used to predict the R/IR (relevant/irrelevant) label using SVM for test data. The predicted labels for 5

 Table 5: Cross-Validation Accuracy for classifying

 Relevant/Irelevant tags

Query	Embedding	Keyword	Embedding & Keyword
Skin Cancer	65.79	66.57	66.28
E-cigarette	69.49	70.94	71.43
MMR-Vaccine	79.98	80.31	84.94
Vitamin-C	80.58	75.18	81.65
Women-HRT	83.74	83.74	82.93

 Table 6: Cross-Validation Accuracy for classifying

 Support/Oppose/Neutral tags

Query	Embedding	Keyword	Embedding & Keyword
Skin Cancer	44.57	54.55	56.6
E-cigarette	47.22	51.33	58.94
MMR-Vaccine	55.3	57.53	62.55
Vitamin-C	55.04	51.8	52.52
Women-HRT	54.88	55.28	52.78

queries are used in further classification of sentences. The system is subjected to 10-fold cross validation while training. The cross validation accuracy obtained from this task using three different approaches - Keyword, Word embedding, Keyword combined with Word embedding respectively is tabulated in Table 5.

Considering the second task, keyword features differs in this case because the keywords contributing R/IR label is different from S/O/N label. So, keywords for further classification are selected based on frequency of occurrence of keywords in support, oppose and neutral statements of training data. Therefore, to classify the sentences into Supporting, Opposing or Neutral, combined feature set which includes the embedding features, keyword features (S/O/N), labels of support, oppose, neutral of training data and predicted R/IR labels taken from task 1 are used for SVM training. The system is subjected to 10-fold cross validation while training. Training results in 5 models for 5 queries that is used for S/O/N (Support/Oppose/Neutral) classifi-

Table 7: Accuracy obtained for Task 1 and Task 2 (in %)

Query	Task 1	Task 2
Skin Cancer	48.8636	23.8636
E-cigarette	76.5625	39.0625
MMR-Vaccine	88.8889	34.7222
Vitamin-C	60.8108	32.4324
Women-HRT	75.8621	43.1034
Overall Accuracy	70.1976	34.6368

cation. SVM predicts S/O/N label for test data. Table 6 tabulates the cross-validation accuracy obtained for the second task using three different approaches -Word embedding, Keyword, Keyword combined with Word embedding respectively. From the cross validation results, it is evident that the method of combination of keyword features and word embedding features is acceptable.

Table 8: Task 1 results by organizers

Team Name	Accuracy	Position
SSN_NLP	78.10	Т
Fermi	77.04	1
JU_KS_Group	73.39	П
Techie Challangers	73.03	11
Jainisha Sankhavara	70.28	Ш
Amrita_CEN	70.19	111

Table 9: Tas	k 2 results	s by organizers	

Team Name	Accuracy	Position
JNTUH	55.43	Т
Fermi	54.87	1
Hua Yang	53.98	II
Techie Challangers	52.46	III
Amrita_fire_CEN	38.53	IV
Jainisha Sankhavara	37.95	V
Amrita_CEN	34.63	VI

Results by CHIS task organizers for our proposed system is tabulated in Table 7.

The accuracy given by CHIS organizers for submission of top 6 teams for task 1 is tabulated in Table 8. Results by organizers for submission of top 7 teams for task 2 is tabulated in Table 9.

6. CONCLUSIONS

In this paper, we have proposed a methodology based on the combination of keyword and word embedding features. These features contribute in the effective retrieval of relevant information. Keyword features for any set of document can be extracted based on its frequency of occurrence. The proposed system will be helpful in extracting the most relevant document for a query, among a large pool of documents in web. Irrespective of the position we have acquired in task 1, our accuracy value is comparable to that of others. The second task is more challenging due to further classification. By considering sentimental features for sentences, accuracy can be increased.

7. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the organizers of Forum for Information Retrieval Evaluation 2016 for organizing a task with great scope of for research. We would also like to thank Xerox Research Centre, for organizing the CHIS task.

8. REFERENCES

 N. Abinaya, N. John, H. Barathi Ganesh, M. Anand Kumar, and K. P. Soman. AMRITA-CEN@FIRE-2014: Named entity recognition for Indian languages using rich features. *ACM International Conference Proceeding Series*, 05-07-Dec-2014:103–111, 2014.

- [2] M. Anand Kumar, S. Shriya, and K. P. Soman. AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages. *CEUR Workshop Proceedings*, 1587:85–88, 2015.
- [3] M. Arun Kumar and M. Gopal. A comparison study on multiple binary-class svm methods for unilabel text categorization. *Pattern Recognition Letters*, 31(11):1437–1444, 2010.
- [4] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 2011.
- [5] L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions when Reading Clinical Reports. *CLEF 2013 Online Working Notes*, 8138, 2013.
- [6] Q. Le and T. Mikolov. Distributed representations of sentences and documents. volume 4, pages 2931–2939, 2014.
- [7] 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.
- [8] G. Remmiya Devi, P. V. Veena, M. Anand Kumar, and K. P. Soman. Entity Extraction for Malayalam Social Media Text using Structured Skip-gram based Embedding Features from Unlabeled Data. *Proceedia Computer Science*, 93:547–553, 2016.
- [9] S. Sanjay, M. Anand Kumar, and K. P. Soman. AMRITA-CEN-NLP@FIRE 2015:CRF based named entity extraction for Twitter microposts. *CEUR Workshop Proceedings*, 1587:96–99, 2015.
- [10] S. Z. Shariff, S. A. Bejaimal, J. M. Sontrop, A. V. Iansavichus, R. B. Haynes, M. A. Weir, and A. X. Garg. Retrieving clinical evidence: a comparison of Pubmed and Google Scholar for quick clinical searches. *Journal of medical Internet research*, 15(8):e164, 2013.
- [11] M. Sinha, S. Mannarswamy, and S. Roy. CHIS@FIRE: Overview of the CHIS Track on Consumer Health Information Search. In Working notes of FIRE 2016 -Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings. CEUR-WS.org, 2016.
- [12] N. Tracy Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. pages 180–188, 2012.