MedSpecSearch: Medical Specialty Search

Mehmet Uluç Şahin[1], Eren Balatkan[1], Cihan Eran[1], Engin Zeydan[2], and Reyyan Yeniterzi[1]

[1] Özyeğin University, İstanbul, Turkey
[2] Türk Telekom Labs, İstanbul, Turkey

Abstract. MedSpecSearch (`www.medspecsearch.com`) is a search engine for helping users to find the relevant medical specialty for a doctor visit based on users' description of symptoms. This system is useful for users who are not sure of which medical specialty they should consult to. Furthermore, the API of the search engine can be used as part of the online doctor appointment and medical consultation sites to route the patient or question to the right medical specialty. The system returns the top 3 relevant specialties when the estimated confidence score is high. Otherwise it asks users to input more data.

Keywords: Text Classification · Word Embeddings · Confidence Estimation · Data Collection

1 Introduction

A recent survey[3] on physician appointment wait times performed over 15 major metropolitan cities in the United States revealed that the time to schedule an appointment has reached an average of 24 days. In addition to the longer wait times, the cost of doctor appointments are getting more expensive worldwide[4]. Given these circumstances, getting appointment from a doctor on the right medical specialty is becoming more crucial for patients not to delay the diagnosis and treatment process any further and spend more. This becomes a more significant problem in places where patients can take their doctor appointments from any medical department without any guidance. For such conditions, the proposed MedSpecSearch system aims to help patients on finding the right medical specialty to visit, by providing a publicly available and user friendly medical specialty search engine. Given the user's description of the medical case like patient information (gender, age etc) and observed symptoms, the MedSpecSearch will return the top 3 medical specialties that can be relevant to the medical case only if the estimated confidence of the returned results are good enough. In case the

[3] https://www.merritthawkins.com/news-and-insights/thought-leadership/survey/survey-of-physician-appointment-wait-times/

[4] https://www.thestar.com.my/news/nation/2018/08/27/doctor-visits-may-cost-at-least-three-times-as-much-next-year/
https://www.thelocal.fr/20170915/some-doctors-visits-are-about-to-get-more-expensive-in-france

user's description of the medical condition is not enough for a confident prediction, the system asks for more information from the user. MedSpecSearch comes with an API which can be used by online physician appointment sites to direct the patient to the right medical specialty for appointment or by online medical question answer sites to route user's question to the right medical specialist.

2 System Description

The Association of American Medical Colleges (AAMC) defined around 30 general and 100 sub-medical specialties[5]. For instance, *Internal Medicine* is categorized as a general specialty, while *Hematology*, *Rheumatology* and *Pulmonary Disease* etc. are categorized as its sub-specialties. The proposed MedSpecSearch system initially focuses on the general specialties and for a given patient description aims to identify the relevant medical specialty within the AAMC categories. The system pipeline is shown in Figure 1 and described in the following sections.

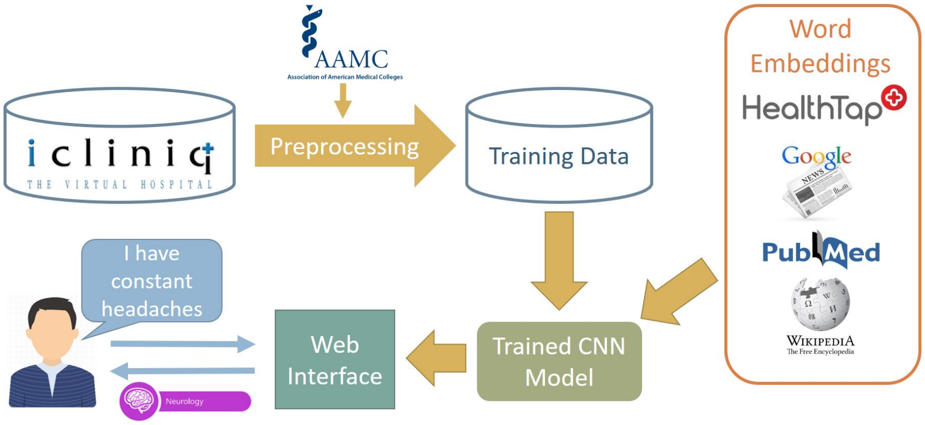


Fig. 1. MedSpecSearch Pipeline

2.1 Data

There are some publicly available de-identified health record datasets (such as the MIMIC [2]) which contain a preliminary, free text diagnosis for the patient on hospital admission. These informative diagnoses are usually generated by the admitting clinicians after listening the complaints and medical histories of patients. These are very useful for routing patients to the right specialty within the hospital; however, due to containing too much medical terminology which general public do not use or even know that existed, they are not useful for

[5] https://www.aamc.org/cim/specialty/exploreoptions/list/

our task which use only the context generated by patients without any medical expert assistance. Therefore, this proposed system explores patient generated content collected from online medical platforms.

Data is collected from two online medical consultation platforms; iCliniq[6] and HealthTap[7]. In iCliniq platform, patients provide details about their conditions and complaints, such as their gender, age, previous critical health history if relevant, previous and current medications, and their symptoms in their questions. Later on, these questions are categorized into a specialty and replied by corresponding medical specialists. We use initial contexts of questions provided by patients as our training input data. The iCliniq has around 90 categories. These categories are not directly used but instead matched to the predefined specialty categories by using the AAMC hierarchy. For instance, questions under *Endocrinology* are categorized under *Internal Medicine* category. In order to check the accuracy of iCliniq category labels, 73 iCliniq questions are randomly chosen and given to two medical doctors for labeling. The inter-rater agreement is calculated with Cohen's [1] kappa statistics for both doctors and doctor-iCliniq pairs. Agreement score of doctors' among themselves is 0.77 which is substantial as being within the 0.6-0.8 range [4]. The agreement scores between iCliniq and doctors are 0.62 and 0.67 which are also substantial. These agreement scores indicate that iCliniq labels are in good quality for supervised learning, hence more than 7K iCliniq questions are retrieved, pre-processed and used for training.

HealthTap is a similar platform with much more questions. Around 1.6 million questions are retrieved from HealthTap, but unfortunately HealthTap uses more diverse set of labels to categorize their questions. It has 230 main categories and around 5400 subcategories. Some example main categories are *abdominal pain*, *ankle* or *blood* which cannot be easily mapped to the AAMC categories. Therefore, HealthTap data is not used in supervised learning but instead explored in an unsupervised manner by training word embeddings.

2.2 The Model

Convolutional Neural Networks (CNN) has achieved competitive performance on many NLP tasks, especially on text classification tasks due to their capability of capturing useful n-grams. Our classification model follows Kim's [3] CNN architecture; however, leverages several pre-trained word embeddings. In order to analyze the individual effects of these embeddings, we use a single channel model as only one embedding is input to the model at once. In addition to our trained Word2Vec model of HealthTap, additional publicly available pretrained word embeddings like Word2Vec [5] embeddings trained with Google News[8] and GloVe [6] embeddings trained on Wikipedia 2014 + Gigaword 5[9] are used. Furthermore, the Word2Vec embedding[10] which is trained by Pyysalo et al.

[6] https://www.icliniq.com

[7] https://www.healthtap.com

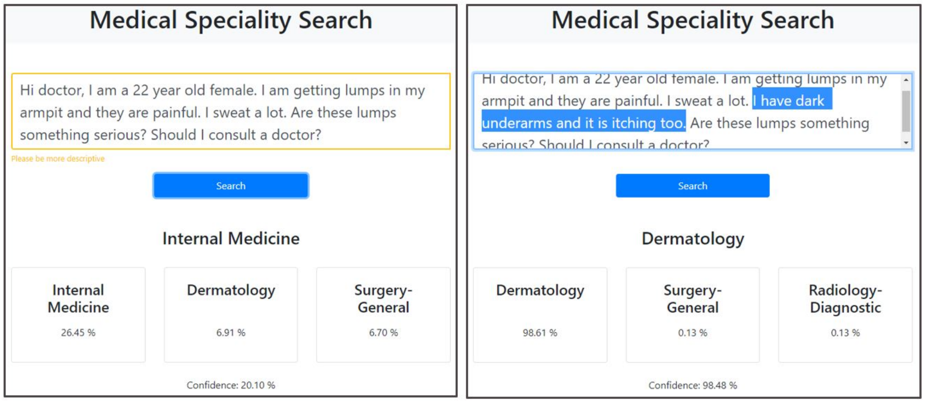
[8] https://code.google.com/archive/p/word2vec/

[9] http://nlp.stanford.edu/data/glove.6B.zip

[10] http://bio.nlplab.org/

[7] on 22 Million PubMed records, 672K PubMed Central Open-Access texts and a recent Wikipedia dump is also used. A 0.8/0.2 split of data to test the performance of these word embeddings returned very similar accuracies: PubMed W2V 72.5%, HealthTap W2V 73.0%, Google News W2V 73.5% and Glove 74.5%. In the advanced options part of the search engine, users can play around with these embeddings by selecting any one them.

If doctors cannot be sure of a diagnosis, they ask for more tests. Even though MedSpecSearch is not a diagnostic system, it uses a similar idea (asking for more user input) when predicting the medical specialty. Being aware of the cost of wrong predictions, the system uses a confidence threshold to decide to either return the ranked list of specialties or not. If the estimated confidence is below the threshold, the system asks user to input more context for describing the medical situation. The default confidence threshold value is 90% but it can be set to any value between 10%-90% in advanced options part.



Hi doctor, I am a 22 year old female. I am getting lumps in my armpit and they are painful. I sweat a lot. Are these lumps something serious? Should I consult a doctor?

Please be more descriptive

Internal Medicine

Internal Medicine	Dermatology	Surgery-General
26.45 %	6.91 %	6.70 %

Confidence: 20.10 %

Hi doctor, I am a 22 year old female. I am getting lumps in my armpit and they are painful. I sweat a lot. I have dark underarms and it is itching too. Are these lumps something serious? Should I consult a doctor?

Dermatology

Dermatology	Surgery-General	Radiology-Diagnostic
98.61 %	0.13 %	0.13 %

Confidence: 98.48 %

Fig. 2. MedSpecSearch Front End with Two Example Queries

An example is provided in Figure 2, where the system's confidence threshold is specifically kept low to show the results of the provided user input. This is an example question from iCliniq. The question on the right is the original question with label *Dermatology* predicted correctly by MedSpecSearch with confidence 98.48%. On the left is the same question with one sentence removed. With this incomplete question the system cannot make a confident prediction, therefore asking for more information from the user is the right call. Using a confidence threshold like 90%, significantly reduces the amount of misclassifications of the system and increases the general prediction accuracy to 90.4%. In our model, we have used the modification proposed by Sensoy et al.[8] to calculate these confidence estimates. All these functionalities are available at `www.MedSpecSearch.com` and the API as well. The API and the trained HealthTap word embeddings can be downloaded from `https://github.com/OzU-NLP/MedSpecSearch`.

References

1. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
2. Johnson, A., Pollard, T., Shen, L., Li-wei, H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., Mark, R.: Mimic-iii, a freely accessible critical care database. Scientific data **3** (2016)
3. Kim, Y.: Convolutional Neural Networks for Sentence Classification. ArXiv e-prints (2014)
4. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ArXiv e-prints (2013)
6. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
7. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine (2013)
8. Sensoy, M., Kandemir, M., Kaplan, L.: Evidential deep learning to quantify classification uncertainty. arXiv preprint arXiv:1806.01768 (2018)