Interactive Symptom Elicitation for Diagnostic Information Retrieval

Tuukka Ruotsalo and Antti Lipsanen
Department of Computer Science, University of Helsinki, Finland
first.last@helsinki.fi

ABSTRACT

Medical information retrieval suffers from a dual problem: users struggle in describing what they are experiencing from a medical perspective and the search engine is struggling in retrieving the information exactly matching what users are experiencing. We demonstrate *interactive symptom elicitation* for diagnostic information retrieval. Interactive symptom elicitation builds a model from the user's initial description of the symptoms and interactively elicitates new information about symptoms by posing questions of related, but uncertain, symptoms for the user. As a result, the system interactively learns the estimates of symptoms while controlling the uncertainties related to the diagnostic process. The learned model is then used to rank the associated diagnoses that the user might be experiencing. Our preliminary experimental results show that interactive symptom elicitation can significantly improve user's capability to describe their symptoms, increase the confidence of the model, and enable effective diagnostic information retrieval.

KEYWORDS

Symptom elicitation; Medical information retrieval

ACM Reference Format:
Tuukka Ruotsalo and Antti Lipsanen. 2018. Interactive Symptom Elicitation for Diagnostic Information Retrieval. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3209978.3210172

1 INTRODUCTION

Amongst the general public, electronic sources are among the most important sources of health and medical information [2]. The ability to find relevant, informative results can be critical in determining whether people seek treatment, or whether a potential diagnosis is found [9]. Despite the increasing availability of medical information on-line, utilizing it for medical advice comes with a risk of experiencing "cyberchondria"; search engines having the potential to escalate medical concerns [10]. To put it simply, searchers are biased to continue searching for more serious diagnosis and search engines are biased to promote popular content so that users are driven to results that other users have preferred instead of content that would be appropriate from a medical perspective.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210172

Recently researchers have developed techniques and gained understanding on how to help laypeople identify credible information [7], but traditional information retrieval techniques applied to medical information rely heavily on ranking functions that are based on learning symptom-disease connections via occurrence statistics [5], and the key focus has been in understanding the semantic variance in user input or query expansion from conventional digital library or Web document sources [3, 11]. Another line of research has focused on diagnostic decision support systems or symptom checkers [4, 8]. However, diagnostic decision support often relies on the assumption that full information about the patient is available and the system is primarily intended to be used by professionals to support clinical decision processes.

Less attention has been devoted to understand the kind of data, inference mechanisms, and interactions with layman users who seek medical advice. For example, consider a layman user suffering from headache and turning into a search engine to find out about possible causes of "headache". The user may receive a variety of information about medical conditions associated with the symptom "headache", some being associated with rare conditions that often also require other symptoms to be present. Moreover, the user may click pages that associate headache with these rare conditions, say brain tumors, and this click data is then used by the search engine to reinforce these pages to rank higher in subsequent searches of other users. As a result, search engines have the potential to increase the anxieties of people who have little or no medical training, instead of providing accurate, reputable, and useful information.

We propose interactive symptom elicitation for diagnostic information retrieval. This approach requires only minimal initial input from the user (e.g. typing in "headache") and interactively elicitates information about related symptoms from the user based on a model learned from electronic health records (e.g. asking about whether the user is suffering from "neck stiffness", "vomiting", "vision defect", "sore throat", or "fever") in order to make a principled inference about the possible diagnosis. The approach can help the user to seek a more objective view of the experienced symptoms and increase the confidence of the information presented for the user. Our approach is in contrast to the previous work as we use reliable data describing symptoms and the related medical conditions estimated from clinical cases. We also tackle the problem of eliciting the correct symptom information from the user by interactively eliciting interaction dialogue between the search system and the user.

We demonstrate a functional system implementing interactive symptom elicitation and report results from a user experiment where layman used the system to detect diagnosis. The results demonstrate the effectiveness and reliability of the estimation of user's medical conditions compared to a conventional typed query driven interaction.

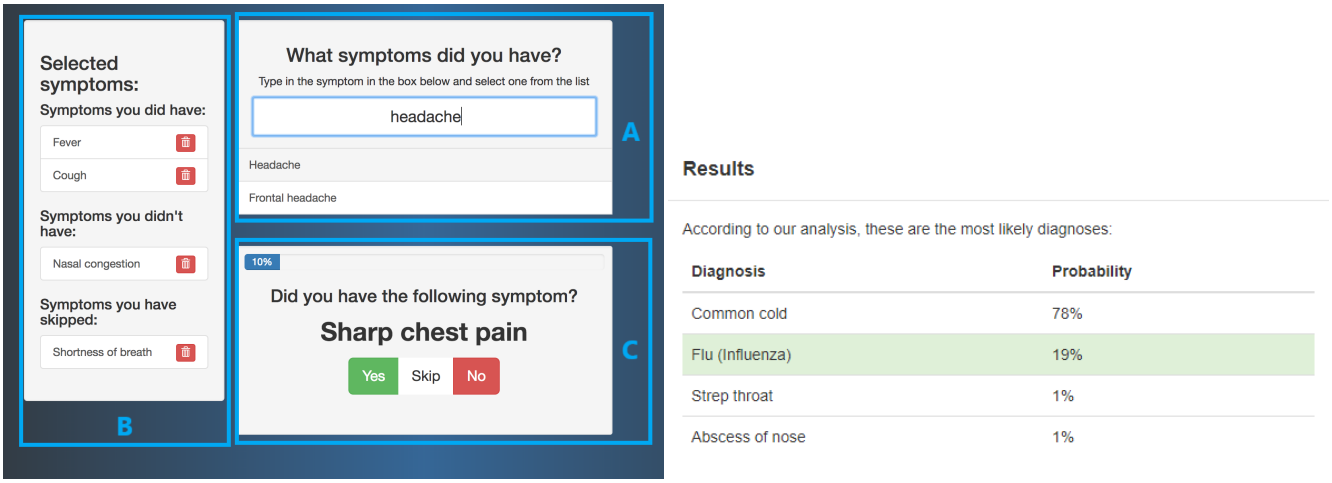

Figure 1: Left: A screenshot of the user interface. The user can input symptoms by typing them into an autocomplete query field (A). The elicited information is shown for the user (B). The system elicitates interaction with the user by asking about symptoms. The user can confirm, reject, or skip a question (C). Right: Results after five iterations. The system ranks common cold and influenza as the two most likely diagnoses. In this case, the influenza (marked with green background) was the correct diagnosis confirmed by the patient.

2 INTERACTIVE SYMPTOM ELICITATION

Interactive symptom elicitation is based on Bayesian optimization by rigorously controlling the uncertainty of the user input. The user can reward or penalize the proposed symptoms and new estimates are learned on-line in response to user interactions [6]. The following subsections describe the user interface, interaction design, and the prediction and ranking models.

2.1 User Interface and Interaction Design

Figure 1 shows the user interface of the system. The user starts the retrieval process by typing symptoms into an autocomplete query field. The system elicitates interaction with the user by asking about related symptoms[1].

At each iteration, a new question about a potential symptom is asked based on the feedback obtained in previous iterations. The user can confirm, reject, or skip a question. In response, the system exploits the feedback elicited from the user (strengthen the most likely diagnosis), but at the same time balances it with exploration (acquire information about related, but yet uncertain symptoms). This learning procedure is called the exploration/exploitation tradeoff of reinforcement learning [1].

Figure 2 shows an example of an interaction sequence and corresponding results. The user has common cold and the user has inserted a symptom "headache", but does not yet have "fever", "nausea", "vision problems", or "seizures". The highest expectation with this input would be to ask the user about "difficulty of swallowing" which is related to some specific diagnoses, such as "Mononucleosis", "Reflux", or "Stroke". However, due to the exploration effect, the system first asks about "cough" which is related to more common respiratory infections and therefore having high uncertainty. The user confirms "cough" and the system subsequentially asks about "nasal congestion", which now has both high expectation, but also high uncertainty. As shown in Figure 2, these questions are asked not because of the highest expectation, but because of the highest upper confidence bound. In other words, "cough" is not the most likely symptom given the user's interaction history, but it has the best potential to provide useful information. After these interactions, the system ranks highest the correct "common cold" diagnosis despite the user did not initially enter the most typical symptoms related to common cold.

2.2 Prediction and Ranking Models

We approach the problem via interactive Bayesian optimization via multi-armed bandits, a type of reinforcement learning, and utilize the upper confidence bound algorithm [1]. In the model, each symptom is an arm of the multi-armed bandit model, and each arm is modeled using a context vector containing all diagnoses associated with the symptom. The algorithm controls the exploration/exploitation tradeoff of the estimates: it predicts expected symptom relevances (exploiting symptoms that have the best fit given the user feedback) and corresponding upper confidence bounds. The symptom with the highest upper confidence bound is selected for interaction. The symptoms with high upper confidence bounds are the ones that have a good fit with less uncertainty (e.g. sore throat for common cold), or lower fit but with greater uncertainty (e.g. neck stiffness for meningitis). These are the symptoms that are optimal for user feedback in order to improve the estimates in the subsequent iterations and avoid only strengthening the most likely estimates.

[1]A video illustration of the system functionality is available at: https://youtu.be/M5etGqiVSFU

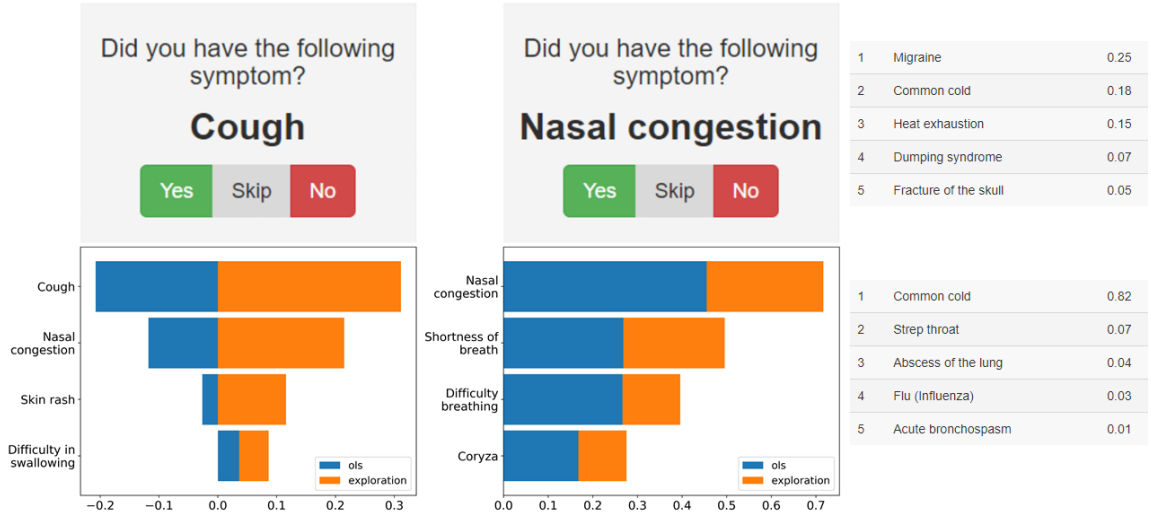

Figure 2: An illustration of an interaction sequence. The left upper panels show the questions for symptom elicitation. The left lower panels show the estimates with the expected value (blue) and upper confidence bound (orange). The upper right panel shows the results before the elicited symptoms and the lower right panel shows the results after the elicited symptoms.

If the system would simply estimate the next symptom to ask by selecting the highest exploitative estimates, the system could suggest symptoms similar to the ones fitting to the present estimate of the diagnosis. This would risk getting stuck in a suboptimal estimate. For example, for the initial symptoms "fever" and "headache", the system would only ask about symptoms fitting to the most common "common cold" diagnosis and not symptoms with a less good present estimate, but the ones that could differentiate the diagnosis. For example, exploring symptoms, such as "neck stiffness" and "dizziness" the system could learn that the user may be suffering from "meningitis" instead of common cold. The selected symptoms are then used to rank the diagnosis using maximum likelihood estimation with Laplace smoothing [12], which is also used to compute the confidence of the estimate.

3 USER EXPERIMENT

A preliminary user experiment was conducted to study the effectiveness of the system in ranking the correct diagnosis in response to user interactions.

3.1 User Experiment

Experimental design A between subjects design was used: each participant used the system only once and with only one diagnosis.
Participants Twelve participants were recruited by word of mouth. The mean age of the participants was 28.33 years. Eight were female, four were male.
Procedure Participants accessed the system over the Internet. Prior to the actual experiment, they read instructions and were asked to insert a medical diagnosis confirmed by a doctor that they had suffered during the previous six months, along with background information, such as age and gender. The entered diagnosis was the target diagnosis used to measure the effectiveness of the ranking. The participants were then requested to inserted initial symptoms as a query that were associated with the diagnosis they entered. After this, the symptom elicitation started. The participants answered a sequence of 20 questions posed by the system. The experiment lasted about 10 minutes. Participation was voluntary and the participants could quit the experiment at any point.
Measures We used four measures to quantify the performance of the system. The mean reciprocal rank, mean rank, and median rank of the target diagnosis were used to quantify the ranking quality. These measures can be used to evaluate systems that return a ranked list of answers when only one of the answers is correct. For example, in our case the system returns a rank of all diagnoses, but only the target diagnosis is the correct diagnosis. In addition, we quantified the confidence of the correct diagnosis. This was measured as the share of the probability mass of the symptoms that the user selected and that were associated with the correct diagnosis.
Data A high-quality databased from Symcat[2] database was used. The data originates from real patient cases describing the symptoms and the associated diagnoses. It consists of 800 diagnoses and 377 associated symptoms. Each symptom is associated with the conditional probability given a diagnosis. Intuitively, it determines the share of patients having a particular symptom when diagnosed with a particular diagnosis.

3.2 Results

The overall results of the user experiment are shown in Table 1. Interactive symptom elicitation shows decreased performance in the first five iterations. The median rank of the correct diagnosis at iteration five is 11.5 compared to the typed query performance of 4.5 at the first iteration. Similar behavior is observed by using Mean Reciprocal Rank and mean rank of the correct diagnosis. The

[2]http://www.symcat.com/

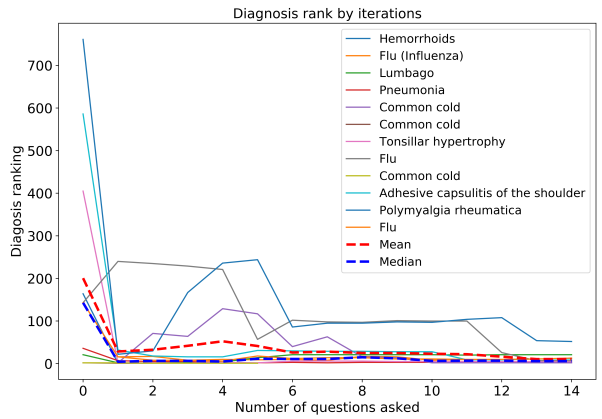

Figure 3: The median rank for all participants and diagnoses over iterations. The dashed blue line shows the mean median rank and the dashed red line the mean rank of the correct diagnosis over all participants.

improvement in ranking, however, was not found to be significant with this sample size (Wilcoxon test, p=0.35). The confidence of the model increases as a function of iterations; the system elicitates information about symptoms that can strengthen the reliability of the estimate of the diagnosis. The explanation for the reduced performance in the first iterations is that the system explores the symptom space. The results show remarkable improvement in the confidence of the estimates from 0.13 of the baseline to 0.65 of the interactive symptom elicitation at 20 iterations (Wilcoxon test, p=0.0005). This shows that by elicitating information about the symptoms the system is able to explore alternative diagnosis, differentiate false diagnosis from the correct or potentially correct diagnosis, and increase the reliability of the diagnosis.

Figure 3 shows the median rank of the correct diagnosis for all participants, and the mean median rank. The same effect of the system exploring alternative, yet possible diagnosis, can be seen between iterations one and six. After six iterations the system starts to converge towards the correct diagnosis. Even difficult and more rare diagnoses, such as Tonsillar hypertrophy, Hermorrhoids, and Polymyalgia Rheumatica are ranked with reasonable performance (i.e. in top 20), even though such conditions may require examination by physician and more advanced diagnostic procedures. While the study is preliminary and the sample size of the study is not yet sufficient for conclusive evidence, it shows that the system is functional and can be effective in real-life usage.

4 CONCLUSIONS

We introduced *interactive symptom elicitation* and demonstrated the technique as a part of a real-world diagnostic information retrieval system. The approach can help to seek more objective view of the experienced symptoms, ask the correct questions to reinforce or discard information about medical conditions, and rank information about medical conditions more accurately than what is possible with conventional signals obtained from the user during information search sessions. The preliminary evidence from a user experiment suggest that 1) the ranking of the correct diagnosis was improved due to interactive symptom elicitation when compared to only typed query input, and 2) the confidence of the correct diagnoses was increased due to interactive symptom elicitation. Our findings have implications for the design of personalized medical information retrieval systems that can help to avoid "cyberchondria" and offer tailored support for individual searchers on reputable medical information sources.

Questions asked	Iter	MRR	Mean	Median	Confidence
Typed query	1	0.362	28.6	4.5	0.13
ISE	5	0.358	41.4	11.5	0.38
ISE	10	0.363	23.2	6.5	0.48
ISE	15	0.39	10.5	4	0.6
ISE	20	0.519	10.1	2.5	0.65

Table 1: Results from the user experiment comparing typed query input and interactive symptom elicitation (ISE). The measures reported are Mean Reciprocal Rank (MRR), mean rank, median rank, and confidence of the correct diagnosis. The results are reported over iterations.

ACKNOWLEDGMENTS

The research was supported by the Academy of Finland (312274).

REFERENCES

[1] Peter Auer. 2003. Using Confidence Bounds for Exploitation-exploration Tradeoffs. *J. Mach. Learn. Res.* 3 (March 2003), 397–422. http://dl.acm.org/citation.cfm?id=944919.944941

[2] Gunther Eysenbach and Christian Köhler. 2002. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 324, 7337 (2002), 573–577. https://doi.org/10.1136/bmj.324.7337.573

[3] Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. 2008. MedSearch: A Specialized Search Engine for Medical Information Retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 143–152. https://doi.org/10.1145/1458082.1458104

[4] Randolph A. Miller. 2016. *Diagnostic Decision Support Systems.* Springer International Publishing, Cham, 181–208.

[5] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a Health Knowledge Graph from Electronic Medical Records. *Scientific Reports* 7, 1 (2017), 2045–2322.

[6] Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2014. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* 58, 1 (Dec. 2014), 86–92. https://doi.org/10.1145/2656334

[7] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1245–1254. https://doi.org/10.1145/1978942.1979127

[8] Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351 (2015). https://doi.org/10.1136/bmj.h3480 arXiv:http://www.bmj.com/content/351/bmj.h3480.full.pdf

[9] Vladan Starcevic and David Berle. 2013. Cyberchondria: towards a better understanding of excessive health-related Internet use. *Expert Review of Neurotherapeutics* 13, 2 (2013), 205–213.

[10] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Trans. Inf. Syst.* 27, 4, Article 23 (Nov. 2009), 37 pages. https://doi.org/10.1145/1629096.1629101

[11] Qing T. Zeng, Jonathan Crowell, Robert M. Plovnick, Eunjung Kim, Long Ngo, and Emily Dibble. 2006. Assisting Consumer Health Information Retrieval with Query Recommendations. *Journal of the American Medical Informatics Association* 13, 1 (2006), 80–90.

[12] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. https://doi.org/10.1145/984321.984322