Smart-MD: Neural Paragraph Retrieval of Medical Topics

Rudolf Schneider
Beuth University of Applied Sciences
Berlin, Germany
ruschneider@beuth-hochschule.de

Sebastian Arnold
Beuth University of Applied Sciences
Berlin, Germany
sarnold@beuth-hochschule.de

Tom Oberhauser
Beuth University of Applied Sciences
Berlin, Germany
toberhauser@beuth-hochschule.de

Tobias Klatt
Beuth University of Applied Sciences
Berlin, Germany
tklatt@beuth-hochschule.de

Thomas Steffek
Beuth University of Applied Sciences
Berlin, Germany
tsteffek@beuth-hochschule.de

Alexander Löser
Beuth University of Applied Sciences
Berlin, Germany
aloeser@beuth-hochschule.de

ABSTRACT

We demonstrate Smart-MD, an information retrieval system for medical professionals. The system supports topical queries in the form [disease topic], such as ["lyme disease", treatments]. In contrast to document-oriented retrieval systems, Smart-MD retrieves relevant paragraphs and reduces the reading load of a medical doctor drastically. We recognize diseases and topical aspects with a novel paragraph retrieval method based on bidirectional LSTM neural networks. We demonstrate Smart-MD on a dataset that contains 3,469 diseases from the English language part of Wikipedia and 6,876 distinct medical aspects extracted from Wikipedia headlines.

CCS CONCEPTS

• **Applied computing → Health care information systems**; • **Information systems → Information retrieval**;

KEYWORDS

Neural Information Classification; Paragraph Retrieval

ACM Reference Format:
Rudolf Schneider, Sebastian Arnold, Tom Oberhauser, Tobias Klatt, Thomas Steffek, and Alexander Löser. 2018. Smart-MD: Neural Paragraph Retrieval of Medical Topics. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186979

1 INTRODUCTION

Medical doctors, in particular at emergencies, often need to make fast decisions and without studying latest research results from journals thoroughly. In particular less experienced doctors might overlook alternative treatments or therapies and often fall back to potentially less effective standard procedures known from their academic studies. Despite the fact that most queries of doctors are of informational intent [14] [13], standard medical search engines, like PubMed[1], still focus on filtering documents for a key word query. Ideally, a doctor could use an effective search engine for retrieving diverse and potentially unknown results from the latest literature about symptoms, therapies, medications, treatments or other often requested aspects during the anamnesis.

Scenario: Consider the case of a doctor searching for treatments of *Lyme disease*, an infectious disease caused by bacteria of the Borrelia type which is mainly spread by *ticks*. She will study essential articles and will find the transmission of ticks from birds to humans as main cause. While she knows from her academic studies that antibiotics such as *doxycycline* will help most patients, she might oversee that certain patients with cardiac diseases will likely suffer from this treatment and should rather be treated with *ceftriaxone*-based antibiotics. Ideally, the system would retrieve all treatments for Lyme disease and would display an aggregated overview of different treatments, including some paragraphs of text which explain infrequent edge cases.

We demonstrate Smart-MD, an information retrieval system that provides such a functionality for medical professionals. It takes as input diseases and a list of optional topical aspects and returns paragraphs that report about the given diseases in context of the given aspects. Moreover, it recognizes and aggregates important facets in these paragraphs, such as correlating medical terms or topics and provides the user these facets for query refinement. Figure 1 shows a typical result for the query 'lyme treatment'. Given the query (1), the system retrieves two highly relevant paragraphs about treatments from two articles on Lyme disease (4) or on Borrelia. The user is able to refine the query with topical aspects that appear in the context of these documents (2). Next, Smart-MD shows a distribution of treatments (3) and the user can narrow the query to a particular novel and previously unknown treatment. Finally, the user may click on an interesting paragraph to inspect the context of the entire document. Thereby the system highlights the topic of each relevant paragraph (6). In particular with long documents, this fine granularity at paragraph level permits the reader to skip many irrelevant passages.

The remainder of this paper is structured as follows: In Section 2, we give details about the neural network models, in Section 3 we outline a walk-through of the system.

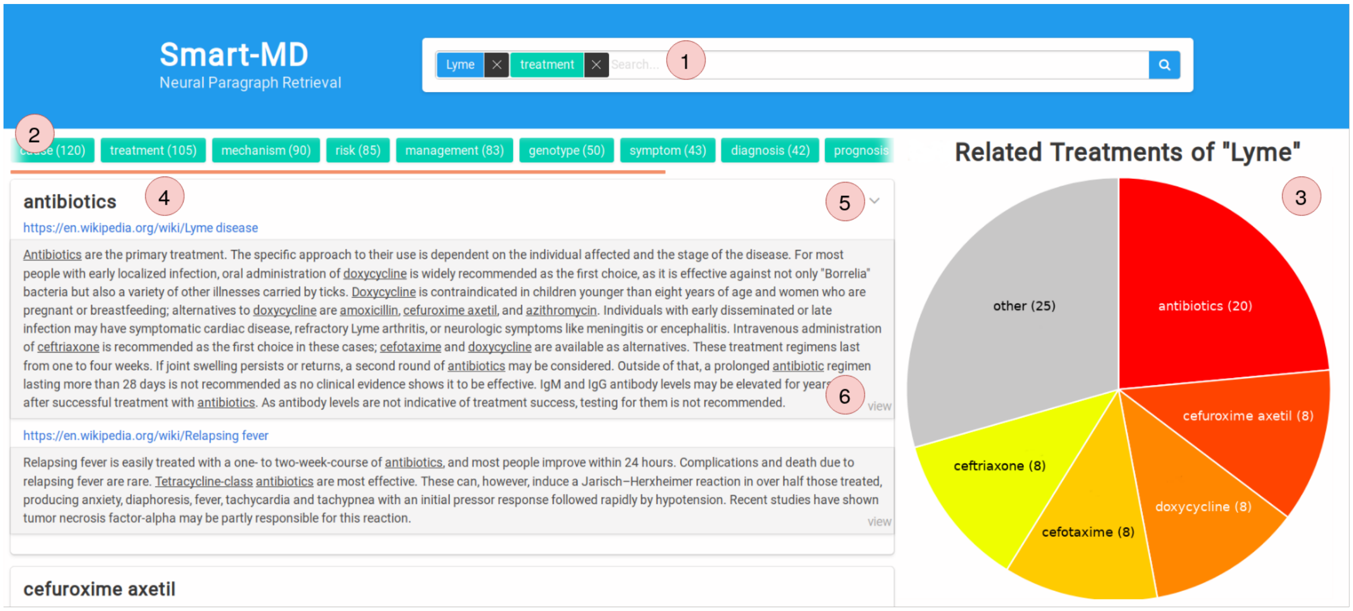
2 PARAGRAPH RETRIEVAL

Smart-MD is built upon two neural information extractors which process the dataset at load time. The *topic extractor* assigns a distribution of topics to each sentence in the dataset. The *entity extractor* recognizes named entities in these sentences. Both models are

[1] https://www.ncbi.nlm.nih.gov/pubmed/

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186979



antibiotics

https://en.wikipedia.org/wiki/Lyme disease

Antibiotics are the primary treatment. The specific approach to their use is dependent on the individual affected and the stage of the disease. For most people with early localized infection, oral administration of doxycycline is widely recommended as the first choice, as it is effective against not only "Borrelia" bacteria but also a variety of other illnesses carried by ticks. Doxycycline is contraindicated in children younger than eight years of age and women who are pregnant or breastfeeding; alternatives to doxycycline are amoxicillin, cefuroxime axetil, and azithromycin. Individuals with early disseminated or late infection may have symptomatic cardiac disease, refractory Lyme arthritis, or neurologic symptoms like meningitis or encephalitis. Intravenous administration of ceftriaxone is recommended as the first choice in these cases; cefotaxime and doxycycline are available as alternatives. These treatment regimens last from one to four weeks. If joint swelling persists or returns, a second round of antibiotics may be considered. Outside of that, a prolonged antibiotic regimen lasting more than 28 days is not recommended as no clinical evidence shows it to be effective. IgM and IgG antibody levels may be elevated for years after successful treatment with antibiotics. As antibody levels are not indicative of treatment success, testing for them is not recommended.

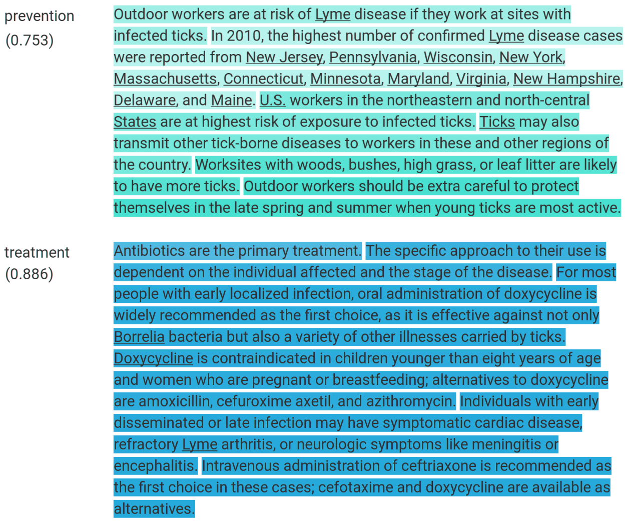
https://en.wikipedia.org/wiki/Relapsing fever

Relapsing fever is easily treated with a one- to two-week-course of antibiotics, and most people improve within 24 hours. Complications and death due to relapsing fever are rare. Tetracycline-class antibiotics are most effective. These can, however, induce a Jarisch–Herxheimer reaction in over half those treated, producing anxiety, diaphoresis, fever, tachycardia and tachypnea with an initial pressor response followed rapidly by hypotension. Recent studies have shown tumor necrosis factor-alpha may be partly responsible for this reaction.

cefuroxime axetil

Figure 1: Screenshot of the Smart-MD user interface. The search bar (1) shows current query terms and offers a auto completion based on the neural entity and topic extractors. The fact distribution chart (3) and the topic tag bar (2) offer visual navigation components to allow the user refinement of the search direction. Smart-MD groups search results by their topics in result cards with a generated title and short description (4). Those can be unfolded using the arrow button (5). The view button (6) opens a full-text view of the document as shown in Figure 2.

trained end-to-end with data from the medical domain, in particular for the diseases scenario. We store all extractions in an index and retrieve them at query time to return relevant paragraphs. In this section we describe these steps briefly.



prevention (0.753)

Outdoor workers are at risk of Lyme disease if they work at sites with infected ticks. In 2010, the highest number of confirmed Lyme disease cases were reported from New Jersey, Pennsylvania, Wisconsin, New York, Massachusetts, Connecticut, Minnesota, Maryland, Virginia, New Hampshire, Delaware, and Maine. U.S. workers in the northeastern and north-central States are at highest risk of exposure to infected ticks. Ticks may also transmit other tick-borne diseases to workers in these and other regions of the country. Worksites with woods, bushes, high grass, or leaf litter are likely to have more ticks. Outdoor workers should be extra careful to protect themselves in the late spring and summer when young ticks are most active.

treatment (0.886)

Antibiotics are the primary treatment. The specific approach to their use is dependent on the individual affected and the stage of the disease. For most people with early localized infection, oral administration of doxycycline is widely recommended as the first choice, as it is effective against not only Borrelia bacteria but also a variety of other illnesses carried by ticks. Doxycycline is contraindicated in children younger than eight years of age and women who are pregnant or breastfeeding; alternatives to doxycycline are amoxicillin, cefuroxime axetil, and azithromycin. Individuals with early disseminated or late infection may have symptomatic cardiac disease, refractory Lyme arthritis, or neurologic symptoms like meningitis or encephalitis. Intravenous administration of ceftriaxone is recommended as the first choice in these cases; cefotaxime and doxycycline are available as alternatives.

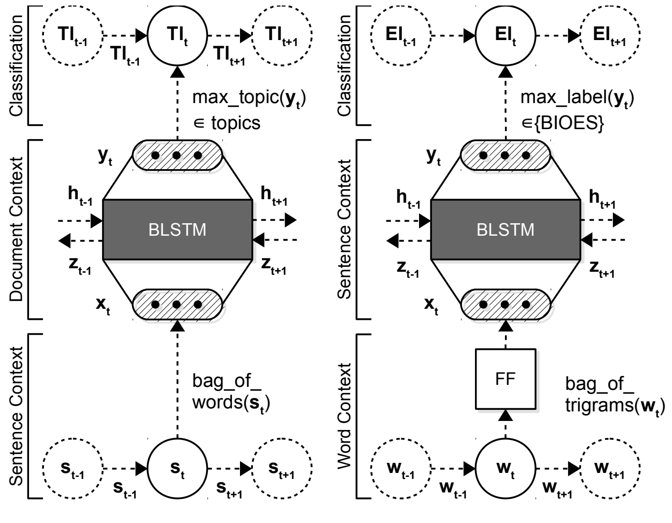
Figure 2: Visualization of the neural topic classification for an example document (excerpt). Smart-MD assigns coherent topic labels 'prevention' and 'treatment' to sentences. The shading of colors visualizes confidence of the best scored class from the prediction, numbers in brackets depict the average confidence per paragraph.

2.1 Sequential Topic Classification

The topic extractor's goal is to assign a coherent distribution of topics over all positions in a document. In contrast to traditional probabilistic topic models such as LDA [3], which describe topic distributions on document-level, we approach to capture topics on sentence level. One possible solution is Paragraph Vectors [9], which treats all paragraphs (or sentences) independently. However, to achieve a coherent sequence of topics, e.g. to spot adjacent sentences that express treatments of a disease, we need to respect the sequential order and long-range dependencies of sentences in the document. Consequential, our approach uses a Long Short-Term Memory (LSTM) network [7] for classification.

Definition of topics from Wiki section headlines. We utilize section and subsection headlines from Wikipedia documents to define possible topics. For example, we observe 6,876 distinct headlines from 3,469 Wikipedia pages on diseases[2]. Table 1 shows the distribution of observed topics among articles. A closer inspection reveals that this distribution is heavily skewed, e.g. top 20 topics cover more than 90% of all paragraphs. We therefore chose 20 representative topic labels for training and assign label 'other' to the remainder. A detailed overview of the topic distribution is shown in Table 2.

[2] taken from the 20170320 Wikipedia dump



(A) Sequential Topic Classification (B) Medical Entity Recognition

Figure 3: Neural network architectures of our section classification model. It classifies a sequence of sentences $\mathbf{s}_t$ to their corresponding section labels $\mathbf{l}_t$. We employ bidirectional LSTMs layers which respect the long-range dependencies of sentences $\mathbf{s}_t$ inside a document.

Sequential classification using BLSTM networks. We utilize the LSTM model with forget gates [5] and bidirectional layers [6] to predict for each sentence s_t a probability distribution y_t for its topic label $\mathbf{Tl}_t = \max(y_t)$. The BLSTM is configured using forward and backward layers with input nodes $\vec{g}_t$, input gates $\vec{i}_t$, forget gate $\vec{f}_t$, output gate $\vec{o}_t$ and internal state $\vec{s}_t$. We encode hidden states $\vec{h}_t$ (forward layer) and $\overleftarrow{z}_t$ (backward layer) for every time step t. We generate the output layer y_t by summing $\vec{h}_t$ and $\overleftarrow{z}_t$.

$$\begin{aligned}
\vec{g}_t &= \phi(\vec{W}_{gx} x_t + \vec{W}_{gh}\vec{h}_{t-1} + \vec{b}_g)\\
\vec{i}_t &= \sigma(\vec{W}_{ix} x_t + \vec{W}_{ih}\vec{h}_{t-1} + \vec{b}_i)\\
\vec{f}_t &= \sigma(\vec{W}_{fx} x_t + \vec{W}_{fh}\vec{h}_{t-1} + \vec{b}_f)\\
\vec{o}_t &= \sigma(\vec{W}_{ox} x_t + \vec{W}_{oh}\vec{h}_{t-1} + \vec{b}_o)\\
\vec{s}_t &= \phi(\vec{g}_t \odot \vec{i}_t + \vec{s}_{t-1} \odot \vec{f}_t)\\
\vec{h}_t &= \vec{s}_t \odot \vec{o}_t \quad / \quad \overleftarrow{z}_t = \overleftarrow{s}_t \odot \overleftarrow{o}_t\\
y_t &= \phi(\vec{W}_{yh}\vec{h}_t + \overleftarrow{W}_{yz}\overleftarrow{z}_t + b_y)
\end{aligned} \tag{1}$$

Our network architecture is shown in Figure 3. We use n-hot bag of words vectors as input features, i.e. $x_t = \sum_{\mathrm{w} \in \mathrm{s}_t} \mathrm{i}_\mathrm{w}$ with indicator $\mathrm{i}_\mathrm{w} \in \{0, 1\}^{|\mathcal{V}_\mathrm{w}|}$ over a fixed vocabulary $\mathcal{V}_\mathrm{w}$. We implement our BLSTM model with 300 cells, sigmoid activation, 0.5 dropout and a softmax output layer. It is trained document-wise with using stochastic gradient descent with ADAM [8], L2 regularization and cross entropy loss using a learning rate of 10^{-3} and backpropagation-through-time [12]. The network classifies a complete document per iteration and is only reset in between documents. We segment the document into paragraphs by splitting at positions where the topic label changes. The outcome of our method is visualized in Figure 2.

2.2 Medical Named Entity Recognition (NER)

The entity extractor's goal is to recognize medical named entities, such as diseases or medications in the documents. This task is often difficult, since for this specialized task only sparse training data exists and recall suffers [10]. We utilize our own work TASTY[3] [1], a generic and robust approach for high-recall named entity recognition and linking in many languages and with sparse training data. TASTY offers strong generalization over domain-specific language, such as in biomedical text (e.g. Medline, PubMed or Wikipedia articles) and can be trained with only few hundred labeled sentences to achieve F1 scores in the range of 84–94% on standard datasets.

Robust recognition using character n-gram embeddings. Similar to the topic extractor, the architecture of our entity extractor utilizes a BLSTM architecture. The model's objective is to assign BIOES entity labels $\mathbf{El}_t = \max(y_t)$ to all words in a sentence [11]. To achieve a robust classifier, we encode words as bag of letter-trigrams as input features, i.e. $x_t = \sum_{\mathrm{tri} \in \mathrm{w}_t} \mathrm{i}_\mathrm{tri}$. This allows us to train a character embedding that is able to recognize typical syllables in a word [2]. We extract possible diseases and other medical entities and store them in the index for query completion and paragraph retrieval.

2.3 Query Processing and Paragraph Scoring

Smart-MD executes queries of the form [disease topic] as follows: First, the user matches ambiguous disease and topic names using the autocomplete. It maps a variety of notations from Wikipedia headlines to well defined classes. We then conduct a conjunctive boolean search and retrieve documents that contain both disease name and topic ID a single document. Finally, we score the candidate paragraphs. Our scoring approach bases on the assumption that paragraphs likely contain medical entities that have a mutual relation with the topic of the paragraph and the requested disease. Moreover, we would like to retrieve for a doctor low frequency events that are probably unknown to the doctor. We measure for each paragraph, proximity between the requested topic and co-occurring entities with normalized pointwise mutual information [4] (nPMI):

$$\mathrm{nPMI}(\mathrm{entity}, \mathrm{topic}) = \frac{\ln \frac{P(\mathrm{entity}, \mathrm{topic})}{P(\mathrm{entity})P(\mathrm{topic})}}{-\ln P(\mathrm{entity}, \mathrm{topic})} \tag{2}$$

$P(\mathrm{entity})$ denotes the probability that retrieved paragraphs contain the entity, $P(\mathrm{topic})$ the probability that the topic is discussed in the retrieved paragraphs and $P(\mathrm{entity}, \mathrm{topic})$ denotes the probability that an entity appears in any retrieved paragraph that discusses the topic. Hence we assign to low frequency events relatively high scores.

3 DEMONSTRATION OUTLINE

We demonstrate Smart-MD in a live demonstration and with a video[4] that shows the case for our query from the introduction ["lyme disease", treatments].

[3] Demo available at http://demo.datexis.com/tasty/

[4] https://www.youtube.com/watch?v=kcDi7qQxpBo

Table 1: Frequency and entropy (H) of top-5 head and randomly selected torso and tail headings for 3,469 diseases and 6,876 distinct headlines in the English Wikipedia.

no.	headline	topic	freq	H
0	Abstract	abstract	3,453	0.03
1	Diagnosis	diagnosis	2,795	0.49
2	Treatment	treatment	2,789	0.49
3	Signs and Symptoms	symptom	1,921	0.69
4	Causes	cause	1,531	0.69
		…		
14	Symptoms	symptom	339	0.32
15	Types	classification	329	0.31
16	Research	research	312	0.30
17	Society and Culture	culture	310	0.30
18	Mechanism	mechanism	224	0.24
		…		
6,873	Fungal Meningitis	other	1	0.00
6,874	Location and Symptoms	symptom	1	0.00
6,875	Molecular Basis of Disease	other	1	0.00

Initial search query. While she is typing the query, the system auto-completes terms against words in the index of diseases or topics. Next, the system retrieves documents, filters, scores and displays top-ranked paragraphs. Now, she can skim the results to get an overview. The system supports her with a short description of the relevant paragraphs of the documents. All sources claiming the same fact are aggregated and can be unfolded by a click on the arrow icon. This representation allows her to overview and skip irrelevant content fast until she reaches interesting treatments.

Query refinement. Smart-MD ranks co-occurring entities and topics in a pie-chart or respectively in the topic bar by their frequency. If resulting paragraphs are still too broad, she can click on topics in the topic bar to refine the query and search for rare facts. Alternatively, she can visit the entity navigation chart on the right that shows a frequency distribution of entities in paragraphs. For accessing less frequent but relevant entities which co-occur with the search query, she clicks on a pie in the chart. This excludes the more frequent entities from the visualization and allows to inspect results in the 'long tail' of search results.

Inspecting the context of a paragraph. Finally, she can drill down into the context of interesting facts by clicking on the text which opens the corresponding document. Next, the system displays the entire document. Similar to hand written notes at margins of a text book, Smart-MD shows an assigned topic for each paragraph. She can now read these pre-labeled topics and skip topics fast until she reaches an important part. She can now drill down further or start over again.

Table 2: Distribution of covered sentences by topics in the wikipedia dump which was used to train the topic extractor. F1 scores are evaluated on a test set of n=32,045 sentences.

topic	freq	F1
abstract	14.35%	82.40
treatment	12.59%	70.55
symptom	11.99%	65.59
diagnosis	11.62%	73.43
cause	10.05%	48.98
other	7.20%	23.17
mechanism	6.28%	58.05
management	4.09%	37.78
epidemiology	4.00%	75.08
history	3.82%	66.19
prognosis	3.08%	62.80
classification	2.29%	37.78
genotype	2.13%	49.60
prevention	1.69%	68.07
culture	1.58%	50.24
research	1.33%	60.09
animal	0.66%	50.25
transmission	0.63%	0.00
risk	0.37%	0.00
complication	0.13%	4.65
screening	0.11%	0.00

3.1 Future Work

We plan in our future work to apply the system to more sophisticated document sources like scientific publications, doctors letters or medical health records.

ACKNOWLEDGEMENTS

Our work is funded by the German Federal Ministry of Economic Affairs and Energy (BMWi) under grant agreement 01MD16011E (Medical Allround-Care Service Solutions), grant agreement 01MD15010B (Smart Data Web) and H2020 ICT-2016-1 grant agreement 732328 (FashionBrain).

REFERENCES

[1] Sebastian Arnold, Robert Dziuba, and Alexander Löser. 2016. TASTY: Interactive Entity Linking As-You-Type. In *COLING'16 Demos*. 111–115.

[2] Sebastian Arnold, Felix A. Gers, Torsten Kilias, and Alexander Löser. 2016. Robust Named Entity Recognition in Idiosyncratic Domains. In *arXiv:1608.06757 [cs.CL]*.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, Vol. 3. 993–1022. Issue Jan.

[4] Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of GSCL*. 31–40.

[5] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. In *Neural Computation*, Vol. 12. 2451–2471.

[6] Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Vol. 385. Springer.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. In *Neural Computation*, Vol. 9. 1735–1780.

[8] Diederik Kingma and Jimmy Ba. 2015. ADAM: A Method for Stochastic Optimization. In *ICLR'15*.

[9] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.. In *ICML'14*, Vol. 32. 1188–1196.

[10] Glen Pink, Joel Nothman, and James R. Curran. 2014. Analysing Recall Loss in Named Entity Slot Filling. In *EMNLP'14*. ACL, 820–830.

[11] Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *CoNLL'09*. ACL, 147–155.

[12] Paul J Werbos. 1990. Backpropagation Through Time: What It Does And How To Do It. In *Proc. IEEE*, Vol. 78. 1550–1560.

[13] Ryen W. White and Eric Horvitz. 2014. From Health Search to Healthcare: Explorations of Intention and Utilization via Query Logs and User Surveys.. In *Journal of the American Medical Informatics Association*, Vol. 21. 49–55.

[14] Illhoi Yoo and Abu Saleh Mohammad Mosa. 2015. Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users. In *JMIR Medical Informatics*, Vol. 3.