Bringing Semantic Structures to User Intent Detection in Online Medical Queries

Chenwei Zhang*¶, Nan Du†, Wei Fan‡‖, Yaliang Li†, Chun-Ta Lu*, Philip S. Yu*§
*Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
†Baidu Research Big Data Lab, Sunnyvale, CA, USA
‡Tencent Medical AI Lab, Palo Alto, CA, USA
§Institute for Data Science, Tsinghua University, Beijing, China
Email: *{czhang99,clu29,psyu}@uic.edu, ‡davidwfan@tencent.com, †{nandu, yaliangli}@baidu.com

***Abstract*—The Internet has revolutionized healthcare by offering medical information ubiquitously to patients via the web search. The healthcare status, complex medical information needs of patients are expressed diversely and implicitly in their medical text queries. Aiming to better capture a focused picture of user's medical-related information search and shed insights on their healthcare information access strategies, it is challenging yet rewarding to detect structured user intentions from their diversely expressed medical text queries. We introduce a graph-based formulation to explore structured concept transitions for effective user intent detection in medical queries, where each node represents a medical concept mention and each directed edge indicates a medical concept transition. A deep model based on multi-task learning is introduced to extract structured semantic transitions from user queries, where the model extracts word-level medical concept mentions as well as sentence-level concept transitions collectively. A customized graph-based mutual transfer loss function is designed to impose explicit constraints and further exploit the contribution of mentioning a medical concept word to the implication of a semantic transition. We observe an 8% relative improvement in AUC and 23% relative reduction in coverage error by comparing the proposed model with the best baseline model for the concept transition inference task on real-world medical text queries.**

***Index Terms*—Intent Detection; Concept Transition; Concept Graph; Neural Network**

1. Introduction

The shortages of healthcare professionals are leading to healthcare systems plagued by bottlenecks. According to the World Health Organization, the world will face a shortfall of nearly 13 million healthcare professionals by 2035 [1]. In the meanwhile, an increasing number of medical-related online services emerge on the Internet to offer ubiquitous medical information to patients via their web search [2]. For example, the Chinese search engine Baidu processes over 6 billion search queries every day, while 60 million of them are healthcare-related text queries[1]. Online medical question answering forums such as xywy.com[2] has more than 22 million unique daily visitors.

With the flourishing demand for medical-related services, it is crucial for service providers to infer implicit user intentions from the diversely expressed medical text queries: what medical concepts a user mentions and how concept transitions are formulated among these concepts. Generally, medical text queries that users search online or post on medical question-answering websites express various medical-related conditions and indicate different information needs, as shown in Table 1.

○ Medical Text Questions ● Inferred Concept Mentions & Concept Transitions
○ Why do I get dizzy so often? ● *Symptom* → *Cause*
○ My three-year-old child is sick with a temperature of 100 degrees she can't keep anything down including liquids. What kind of medicine should I give my child, and how much? ● Symptom → *Medicine* → *Instruction*
○ Do I have insomnia if I have trouble staying asleep? Any medication is recommended to help me fall asleep easier? ● *Disease* ← *Symptom* → *Medicine*

TABLE 1. MEDICAL QUERIES AND THE EXTRACTED MEDICAL CONCEPT MENTIONS &TRANSITIONS. BEST VIEWED IN COLOR.

Usually, medical semantics are formulated by users during their efforts to express their existing medical conditions as well as their intended medical-related information needs, either explicitly or implicitly.

Concept Mention: Queries are expressed diversely by mentioning different types of medical concepts, where a concept c is defined as a group or class of objects and/or abstract ideas representing similar fundamental characteristics in a certain domain. $C = \{c_1, c_2, ..., c_M\}$ is list of a full spectrum of M concepts in a specific domain. For example, in the medical domain, C = {disease, symptom, medicine, ...}. Concepts can be mentioned in a query by specific object names as explicit mentions ("Tylenol", "Ibuprofen"

¶ *Part of the work was done when the author was an intern at Baidu Research Big Data Lab.*
‖ *Part of the work was done when the author was employed by Baidu Research Big Data Lab.*

1. http://science.china.com.cn/2016-1124content_9180719.htm
2. http://club.xywy.com

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

or "xxx caplet/capsule/drop/syrup"), as well as implicit mentions by abstract ideas that refer to the concept (e.g. "remedy") or phrases indicating this concept (e.g. "which medication/medicine/drug").

The way concept mentions organized in a question naturally forms concept transitions that reflect the information-seeking goal of users. Such ubiquitous observation in medical text queries is rarely studied in previous literatures. A typical formulation for medical intent detection is to model each semantic transition as a single label [3], or as a two-element tuple indicating 1) what a user have described and 2) what information the user is looking for [4]. In this work, we define the transition between concepts as:

Concept Transition: A concept transition $t_{i \to j}$ exists in a query when two concepts c_i, $c_j \in C$ are mentioned (either explicitly or implicitly) with a semantic transition between them, where c_i provides contexts and c_j serves as the goal for information seeking.

For example, medical queries with concept transitions $t_{Symptom \to Medicine}$ usually start with patients describing their symptoms and asking for related information about medications that help them alleviate their symptoms. T contains the full spectrum of N concept transitions in a certain domain, which can be indexed as $T = \{t_1, t_2, ..., t_N\}$ for simplicity. Those two index notations are used interchangeably in this paper. We can have $t_{Symptom \to Disease}$ and $t_{Symptom \to Medicine}$ for the last query in Table 1. This formulation defines each tuple as an individual label and fails to consider the interactions among multiple medical concept transitions in a medical query, which prevent them from satisfactorily detect sophisticated user intentions with complex semantic structures. In real-world medical text queries, multiple semantic transitions in a single question may conjugate with each other by mentioning the same medical concept. For example, $t_{Symptom \to Disease}$ and $t_{Symptom \to Medicine}$ share the same concept $Symptom$ by expressing symptoms: "trouble staying asleep" and "fall asleep" in the query.

Alternatively, we can bring semantic structures to concept transitions by formulating multiple concept transitions over a directed concept graph with the following definition:

Concept Graph: Let $G = \langle C, T \rangle$ be a concept graph where each node represents a concept $c_m \in C$ and $t_{i,j} \in T$ be a directed edge from node c_i to c_j. A concept graph G is a graph-based formulation of concepts and concept transitions in a certain domain.

Multiple concept transitions in a query may follow a natural chain-like path, such as the path $Symptom \to Medicine \to Instruction$. Thus for a given query Q, the structured semantics can be represented by the subgraph of the concept graph, namely the active concept graph:

Active Concept Graph: Let an active concept graph $G_Q = \langle C_Q, T_Q \rangle$ be a subgraph of $G = \langle C, T \rangle$, indicating concepts $C_Q \subseteq C$ mentioned in a query Q and concept transitions $T_Q \subseteq T$ activated by the the query Q.

For example, with a graph-based formulation, the concept transition for the second question in Table 1 is formulated as $Symptom \to Medicine \to Instruction$ since the user first describes his/her symptoms ("sick", "temperature of 100 degrees") and inquires about information on the *medicine* concepts ("What kind of medicine"), followed by phrases ("and how much") indicating further information seek intentions about *instructions* on the medicine. Real-world text questions often exhibit a mixture of multiple concept transitions in each question in which shared concept mentions serve as a bridge coupling two or more concept mentions [5]. Thus, a graph-based formulation would essentially allow us to jointly model and infer correlations between concept mentions and multiple concept transitions simultaneously, which is one of our key contributions.

The Concept Transition Inference Problem: In order to better capture a focused picture of people's medical-related information search and information access strategies, we propose and study the concept transition inference problem for online medical queries with a graph-based formulation. Given 1) a text query Q which consists of K elements $\{q_1, q_2, ..., q_K\}$, where each element is a word or a phrase and 2) a concept graph $G = \langle C, T \rangle$, where C denotes concepts and T indicates concept transitions, the concept transition inference problem tries to effectively infer an active concept graph $\hat{G}_Q = \langle \hat{C}_Q, \hat{T}_Q \rangle$ given a query Q. Figure 1 illustrates this idea.

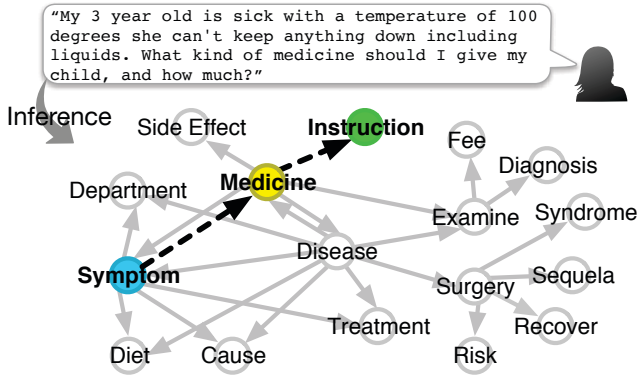

Figure 1. The concept transition inference problem that extracts a directed, subgraph indicating the structured semantics for each query. $\hat{C}_Q$ are shown as colored nodes and $\hat{T}_Q$ are shown as black dashed lines.

Challenges: A typical solution for the concept transition inference problem evolves hand-engineering features based on expert knowledge in the medical domain, such as using pre-defined rules [6] or templates [7], or constructing a word-concept mapping dictionary [4] for question intent classification. Even if one discounts the tedious effort required for feature engineering, those features are usually designed for a limited number of questions accessible to domain experts and do not generalize to handle various user expressions in real-world medical text questions. People with different knowledge background tend to express the same idea in different ways. For example, a medicine concept can be mentioned by specific drug names such as "Tylenol", "Ibuprofen" or phrases like "what kind of medicine/drug/medication". The decent performance of those approaches usually comes at the cost of acquiring an external knowledge base to handle varying linguistic

modalities and diversified expressions. How to minimize feature engineering without compromising the performance for the concept transition inference is still challenging.

Moreover, comparing with general-purpose text questions which people have been posting or searching for online, where users only focus on a single concept (such as "weather", "politics" or "stocks"), concept transitions in medical queries usually involve rich semantics. It would take strenuous efforts to model correlations among multiple concept transitions without considering the shared concept mentions effectively. What's more, unlike many existing works on medical text analysis such as sentiment classification [8], [9] which consider positive, negative or neutral sentiments in medical texts, it is challenging yet rewarding to consider structured concept transitions that model complex medical semantics in real-world medical text queries.

Overall, our paper makes the following contributions:

1) We observe and formally define concept transitions in medical text queries and show appealing properties among multiple concept transitions with shared concept mentions.
2) We study the concept transition inference problem with a graph-based formulation, which brings semantic structures to diversely expressed natural language queries.
3) We propose an end-to-end solution with a novel neural network model to the concept transition inference problem without additional external knowledge requirements.
4) We empirically evaluate the proposed method on real-world medical text queries.

2. Medical Concept Transition Inference

In this section, a novel neural network structure is introduced to provide an end-to-end solution to the concept transition inference problem where the input is a text query and the output is an active concept graph inferred by the given query. The model utilizes word representations to deal with the lexical diversities. Also, part-of-speech embedding of each word is used to further capture the syntax information. Recurrent neural networks are adopted to model the sequential information from distributed representations of word and POS tag sequences in each query simultaneously. In the graph-based co-inference procedure, concepts and concept transition are inferred collectively. A concept encoder is proposed to utilize the joint outputs of two RNNs to encode each element into a concept vector. Especially, the concept encoder is able to learn a confidence score which indicates the contribution of each element in encoding concept mentions in a query. While for inferring concept transitions, a transition encoder learns to summarize the semantics and construct a transition vector, from which we infer a probability distribution on all possible concept transitions. The loss of the neural network structure not only incorporates prediction errors between the predict concept transitions and the true concept transitions but also exploit a mutual transfer loss indicating the conflicts between the inferred concepts and their corresponding concept transitions. An active concept graph is presented with the inferred concepts and concepts transitions, by collectively minimizing a graph-based mutual transfer loss based on the concept graph. Figure 2 gives an overview of the proposed method.

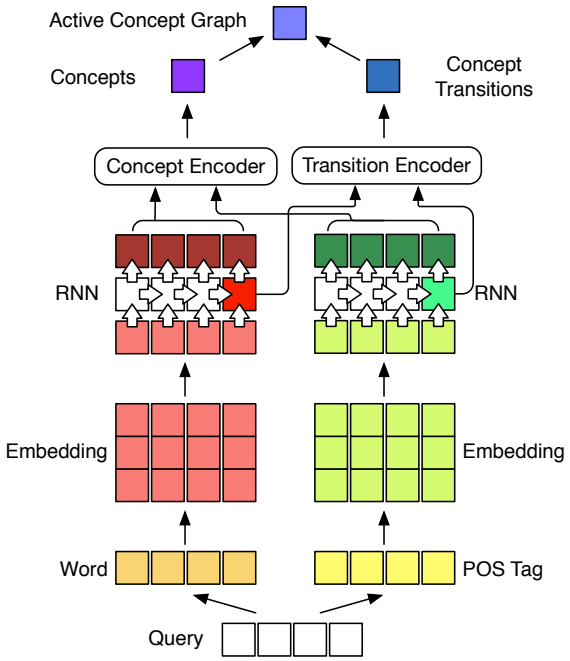

Figure 2. The proposed neural network architecture.

2.1. Lexical-Syntax Representations

Unlike traditional methods which ignore the sequential information of the input text query and treat it as a bag-of-words (BoW), in this work a text query Q is considered as a sequence of elements $\{q_1, q_2, ..., q_K\}$, where each element q_k can be a word or a phrase. K is the length of a text query, which varies from different text queries. For each element q_k in a text query Q, we utilize both word representations indicating the lexical information, as well as its corresponding Part-of-Speech (POS) tag as the syntax information.

Part-of-speech (POS) tags bring useful syntax information about general word categories (such as noun, verb, adjective, etc.), which is especially helpful in dealing with ambiguous words and diversified expressions. For example, "fever" can be either a noun or a verb. The word "fever" with a POS tag "noun" is defined as a disease that causes an increase in body temperature and the fever with a POS tag "verb" can be considered as someone in a fever, as a symptom. In this work, an existing POS tagger[3] is utilized to give general POS tags to each element in the query. The lexical-syntax joint representation consists of words

3. https://github.com/fxsjy/jieba

along with POS tags are shown to be effective in modeling both lexical (words) and syntax (POS tags) from the natural language text corpus in various tasks [10], [4]. In this work, each element Q_k of a query Q is represented by words and POS tags as a tuple:

$$q_k = (w_k, p_k) \ \ s.t.\ w_k \in \mathbb{R}^{V_{word}}, p_k \in \mathbb{R}^{V_{pos}}, \quad (1)$$

where w_k is the one-hot representation of the k-th word in the query Q and V_{word} is the number of unique words, namely the vocabulary size. Similarly, p_k is the one-hot representation of the k-th word's POS tag in the query. V_{POS} is the POS vocabulary size.

2.2. Word Embedding

The one-hot representation suffers from the curse of dimensionality since the representation becomes extremely sparse as the vocabulary becomes large. The word embedding is used to transfer one-hot representation of each word w_k and POS tag p_k into a dense representation:

$$w\_embed_k \in \mathbb{R}^{D_{word}}, p\_embed_k \in \mathbb{R}^{D_{pos}}, \quad (2)$$

where V_{word} usually can be large up to millions while D_{word} is reduced to several hundreds. Note that D_{word} and D_{pos} are usually set empirically. In this work, we set $D_{word} = 100$ and $D_{pos} = 20$. The embedded representation of each w_k and p_k are learned respectively by a linear mapping via a skip-gram model [11]:

$$\begin{aligned} embed\_w_k &= E_{word}\ w_k \\ embed\_p_k &= E_{pos}\ p_k, \end{aligned} \quad (3)$$

where $E_{word} \in \mathbb{R}^{D_{word} \times V_{word}}$ and $E_{pos} \in \mathbb{R}^{D_{pos} \times V_{pos}}$ are weights. The skip-gram learns a distributed representation of each word or POS tag based on its context. In medical text queries, that means an explicit mention of a concept ("Tylenol") and an implicit mention of a concept ("Which medicine") may have similar representations when they occur in similar context, when trained properly. That helps us solve the diversified expressions in medical text queries.

In this work, the embedding is initialized with word vectors pre-trained from 64 million medical text queries and updates with the model during training. After the word embedding, the k-th element in the text query q_k has a lexical-syntax representation, represented by a tuple:

$$e_k = (embed\_w_k, embed\_p_k). \quad (4)$$

2.3. Recurrent Neural Network

Once we obtained a representation e_k for each element q_k in a query Q, the $embed\_w_k$ sequence and the $embed\_p_k$ sequences are fed into two recurrent neural networks, namely $\mathrm{RNN_W}$ and $\mathrm{RNN_P}$, to capture the sequential semantics in the question respectively.

In general, a recurrent neural network keeps hidden states over a sequence of elements and update the hidden state h_k by the current input x_k as well as the previous hidden state h_{k-1} where $k > 1$ by a recurrent function:

$$h_k = \mathrm{RNN}(x_k, h_{k-1}) \quad (5)$$

The Gated Recurrent Unit (GRU) [12] is proposed to address the gradients decay or exploding problem [13], [14] over long sequences in the vanilla RNN. The GRU has been attracting great attention since it overcomes the vanishing gradient in traditional RNNs and is more efficient than LSTM [15] on certain tasks [16]. The GRU is designed to learn from previous time stamps with long time lags of unknown size between important time stamps. Note that, the output vector is not gated in GRU, which makes the GRU more appealing to our problem as no partial view will be given for each output vector o_k.

In this work, two separate RNN with GRU cells, namely RNN_W and RNN_P, are adopted to model the sequential information for the sequence of embedded words $embed\_w_k$ and the sequence of embedded POS tags $embed\_p_k$:

$$\begin{aligned} h\_w_k, o\_w_k &= \mathrm{RNN_W}(embed\_w_k, h\_w_{k-1}) \\ h\_p_k, o\_p_k &= \mathrm{RNN_P}(embed\_p_k, h\_p_{k-1}), \end{aligned} \quad (6)$$

2.4. Graph-based Co-inference

In order to fully exploit the correlations of concepts and corresponding concept transitions, concepts and concept transitions are inferred collectively from a concept graph for each query. The concept inference aims to select a subset of concepts $\hat{C}_Q \subseteq C$ that are mentioned in a query Q, implemented by the concept encoder. To inference transitions, a transition encoder is utilized. The concepts $\hat{C}_Q$ and transitions $\hat{T}_Q$ are inferred collectively, by minimizing a mutual transfer loss which indicates the conflicts within the collectively inferred active concept graph $\hat{G}_Q =< \hat{C}_Q, \hat{T}_Q >$.

2.4.1. Concept Encoder. In concept inference, a concept encoder is proposed to encode all the concept mentions from a sequence of output states of an RNN to concept vectors accordingly. Since some words in a query may contribute more to a concept mention in a query while some other words are less contributive, the concept encoder itself learns to assign a confidence score to each output state. Let o_k be the k-th output vector of an RNN, while in this work we concatenate the output vectors of RNN_W and RNN_P:

$$o_k = [o\_w_k, o\_p_k], o\_w_k \in \mathbb{R}^{1 \times D_{o_w}}, o\_p_k \in \mathbb{R}^{1 \times D_{o_p}}, \quad (7)$$

where D_{o_w} and D_{o_p} are the output dimension of output vectors in RNN_W and RNN_P. The concept encoder assigns a score s_k for each o_k indicating the degree of confidence, parameterized by θ:

$$s_k = CE(o_k; \theta) \quad \text{s.t.} \sum\nolimits_k s_k = 1, \forall s_k \in [0, 1]. \quad (8)$$

All s_k scores in a query are normalized to sum up to one. In this work, the concept encoder is implemented as a single layer neural network with a non-linear activation function ReLU. Thus $\theta = \{W_\theta \in \mathbb{R}^{(D_{o_w}+D_{o_p}) \times 1}, b_\theta \in \mathbb{R}\}$. Note that although weights and biases are applied on each of the o_k, they are shared among all $o_1, o_2, ..., o_k$. Figure 3 shows the architecture of the concept encoder. The

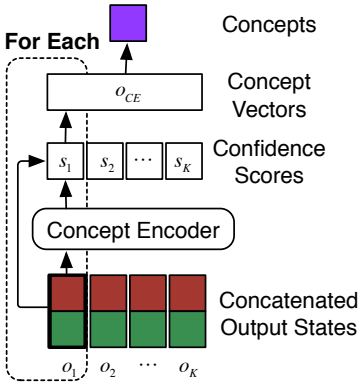

Figure 3. The concept encoder is used to determine confidence scores for each joint output state. This figure shows an example of a score s_1 learned from the concept encoder for o_1.

$o_{CE} \in \mathbb{R}^{(D_{ow}+D_{op})\times K}$ is a representation of encoded concepts from the query, which is calculated as follows:

$$O_{CE} = \left[\ (CE(o_1;\theta)\cdot o_1)^T \quad ... \quad (CE(o_K;\theta)\cdot o_K)^T \ \right] \tag{9}$$

The probability that a concept $c_i \in C$ is activated in a query Q is defined as:

$$(\hat{C}_Q)_m = P(c_m|c_m \in C;\theta) = \frac{1}{1+e^{-W_{CE}O_{CE}+b_{CE}}}, \tag{10}$$

where $W_{CE} \in \mathbb{R}^{1\times(D_{ow}+D_{op})}$, $b_{CE} \in \mathbb{R}$ are weights and biases for such probability inference. We use $\hat{C}_Q \in \mathbb{R}^{1\times M}$ to quantify the probability distribution on all M concepts for a given query Q.

2.4.2. Transition Encoder. In the field of machine translation, a novel recurrent neural network encoder-decoder has gained attention [17], where the encoder recurrent neural network encodes the global information spanning over the whole input sentence in its last hidden state. Inspired by the effectiveness of the last hidden states in modeling natural language sequences in applications like dialog systems [18], we propose a transition encoder which leverages the last hidden state of the neural network for both RNN_W, RNN_P to make inferences on concept transitions, where the transition vector o_{TE} is constructed by $O_{TE} = [h\_w_K, h\_p_K]$, where K is the length of the query. The probability of a transition $t_n \in T$ given the query Q is quantified by:

$$(\hat{T}_Q)_n = P(t_n|t_n \in T;\phi) = \frac{1}{1+e^{-W_{TE}O_{TE}+b_{TE}}}, \tag{11}$$

where $\phi = \{W_{TE} \in \mathbb{R}^{1\times(D_{ow}+D_{op})}, b_{TE} \in \mathbb{R}\}$ parameterizes weights and biases for the transition encoder. Similarly, $\hat{T}_Q \in \mathbb{R}^{1\times N}$ denotes the inferred probability distribution of all N concept transitions given a query Q.

2.5. Mutual Transfer Loss

The idea of mutual transfer loss is to characterize the loss caused by transferring the inferred concept transitions to their corresponding concepts, and the other way around. Since for each concept transition $t_{i\to j} \in T$, two concepts c_i and c_j are evolved in the query. If a concept transition $t_{i\to j}$ is inferred with a high probability while its corresponding concepts c_i, c_j have low probabilities, then that indicates conflicts in the final active concept graph. The mutual transfer loss is proposed in the co-inference procedure to minimize the conflicts between the inferred concepts and concept transitions so that the resulting active concept graph can be more reasonable.

The graph-based formulation for concept graph gives an appealing property that transitions and their proximate concepts can be clearly characterized by a transfer matrix $A \in \mathbb{R}^{M\times N}$ over the concept graph $G = \langle C, T\rangle$. Each entry $a_{mn} = 1$ if and only if the concept c_m involves in at least a concept transition $t_{m\to\cdot}$ or $t_{\cdot\to m}$. The mutual transfer loss is defined on $\hat{C}_Q, \hat{T}_Q, \mathcal{T}_Q$ as:

$$\mathcal{L}_{MTL}(\hat{C}_Q, \hat{T}_Q, \mathcal{T}_Q) = H(\mathcal{T}_Q, \hat{T}_Q) + E(\hat{C}_Q, \hat{T}_Q), \tag{12}$$

where $\mathcal{T}_Q$ is a ground truth one-hot indicator for concept transitions given a query Q. $\hat{C}_Q$ and $\hat{T}_Q$ are inferred concepts and concept transitions with the proposed method. $H(\cdot,\cdot)$ calculates the cross entropy [19]. $E(\hat{C}_Q, \hat{T}_Q)$ is an energy-based function on inferred transitions $\hat{T}_Q$ and inferred concepts $\hat{C}_Q$. Each combination of $\hat{C}_Q$ and $\hat{T}_Q$ corresponds with an energy value, the lower energy level a combination of $\hat{C}_Q$ and $\hat{T}_Q$ has indicates less conflicts among the inferred concepts and transitions. In this work, an energy-based function for $E(\hat{C}_Q, \hat{T}_Q)$ is proposed as:

$$E(\hat{C}_Q, \hat{T}_Q) = \mathcal{L}_R(\hat{C}_Q, \hat{T}_Q A^T) + \mathcal{L}_R(\hat{T}_Q, \hat{C}_Q A), \tag{13}$$

where $\mathcal{L}_R$ is implemented by a ranking loss function [20] that penalizes cases where the inferred concepts/transitions after transformation by matrix A have high probabilities but order below the ranking of the originally inferred concepts/transitions in a query. $\mathcal{L}_R$ has a general form:

$$\mathcal{L}_R(\hat{X}, \hat{Y}) = \frac{1}{\left|\hat{X}\right|\left(L - \left|\hat{X}\right|\right)} \left|\{(p,q) : \hat{Y}_p < \hat{Y}_q, \hat{X}_p \geq \hat{X}_q\}\right|, \tag{14}$$

where $\hat{X} \in \mathbb{R}^{1\times L}$ is the originally inferred labels and $\hat{Y} \in \mathbb{R}^{1\times L}$ is the inferred labels from the transformation with A. $|\cdot|$ denotes the number of ground truth labels being assigned. L is the label size, where we have M for concepts and N for concept transitions.

3. Evaluation

3.1. Data Set

We collect medical queries from an online medical question answering forum[4], on which user posted their healthcare related questions and medical professionals give online suggestions or advice. The obtained corpora are in Chinese. Due to the fact that Chinese text queries are not naturally split by spaces, word segmentation is performed using a Chinese word segmentation package [21].

4. http://club.xywy.com

After preprocessing and annotation, we obtain 10,000 medical text queries. We end up having 17 unique types of concepts and 23 unique types of concept transitions, among which 11,531 unique words and 60 unique POS tags are observed. A medical text query has the following format: { "text": "宫颈管 慢性 炎症 伴 鳞状 上皮 内 挖空 细胞 聚集 是 宫颈癌 吗 严重 吗 需要 leep 手术 吗", "pos": "n b n v n n n n n v v n y a y v eng n y", "concept": "fee|disease|surgery|recover|treatment", "concept_transition": "disease → surgery → recover"}, where the POS tagging uses ICTCLAS annotation [22]. The average length of question is 13.8, with a standard variation of ±6.1. The average number of concepts in labeled queries is 3.6020±0.8. The average number of concept transitions is 2.4723±0.7.

Word embeddings are pre-trained using a skip-gram model [11] on 64 million unlabeled medical text queries separately. Context window size is set to 8 and we specify a minimum occurrence count of 5. The vocabulary contains 100-dimension vectors on 382216 words. Words not presented in the set of pre-trained words are initialized as random vectors. All word vectors will be updated during training.

3.2. Experiment Settings

Comparison Methods: To show the advantages of the proposed method in addressing the concept transition inference problem, we compare it with the following baseline models.

- LR: a logistic regression model applied with POS tagging features and word representations.
- NNID-JM [4]: the neural network intention detection model with joint modeling. Both words and POS tags are used to characterize the question . Domain-specific POS tags, such as "noun_medicine", are used in NNID-JM instead of "noun" for word "Tylenol". The NNID-JM doesn't explicitly exploit label correlations on the output level.
- CI: the concept inference model which only infers mention of concepts from queries with the concept encoder. $H(\mathcal{C}_Q, \hat{C}_Q)$ is used as the loss function for the CI task.
- CTI: the concept transition inference model without co-inference. Only concept transitions are inferred from queries without considering concepts. The last output states of two RNNs are concatenated to predict the concept transitions. $H(\mathcal{T}_Q, \hat{T}_Q)$ is used as the loss function.
- coCTI: the concept transition inference model with co-inference. $H(\mathcal{T}_Q, \hat{T}_Q) + H(\mathcal{C}_Q, \hat{C}_Q)$ is used as the loss function. This variation can be seen as a multi-task learning model for concept and concept transitions, where both tasks share the neural network structure for word representation.
- coCTI-MTL: the proposed model with co-inference and a mutual transfer loss $\mathcal{L}_{MTL}$, where the CI task and CTI task not only share the neural network structure, but also guided by the mutual transfer loss.

Evaluation Metrics: Each edge in the concept graph is considered as an individual label and we evaluate inferred concept transitions as a multi-class, multi-label classification problem. *Receiver operating characteristic (ROC)*, the *micro/macro-average area under the curve (micro-AUC, macro-AUC)*, *coverage error* and *label ranking average precision (LRAP)* are used to evaluate the effectiveness of the proposed model in inferring concept transition in medical text queries.

Experiment Settings: The embeddings for word and POS tagging have a dimension of 100 and 20, respectively. The hidden layer and the output layer of the GRU unit have a dimension of 100. For training the proposed neural network structure, 70% of the labeled data are used for training and 10% data serve as a validation set to tune for the best parameter set. The remaining data are used for testing. Cross-validation is used and we combine test data in each fold to report the test performance. The optimization is performed in a mini-batch fashion with a batch size of 32. The Adam Optimizer [23] is applied to train the neural network and the initial learning rate is set to 10^{-4}. Weight variables are initialized with the Xavier initializer [24] and bias variables are initialized as zeros.

3.3. Evaluation Results

Figure 4 shows the effectiveness of the proposed model by micro-AUC and ROC curves. Generally, neural network based models (NNID-JM, CTI, coCTI, coCTI-MTL) outperform traditional logistic regression model (LR) consistently. For NNID-JM, in order to make a fair comparison, domain

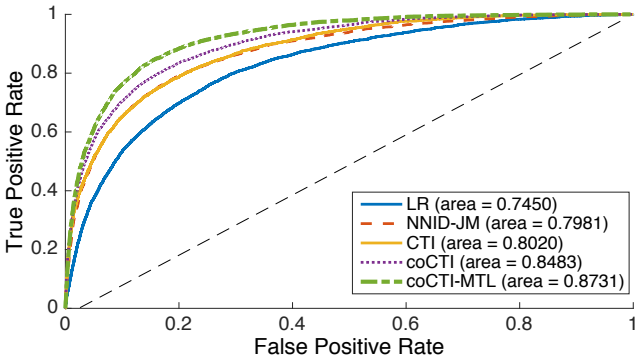

Figure 4. micro-AUC scores and ROC curves.

specific POS tags (such as noun_disease, noun_medicine, noun_symptom) are maintained as an external knowledge base. Those POS tags are used by the POS tagger in NNID-JM as its default setting. When compared with NNID-JM, the proposed CTI model achieves similar performance on micro-AUC, while it doesn't rely on any other external knowledge like domain-specific POS tags in NNID-JM. In practical, utilizing a concept transition graph is usually more feasible than tagging words and building dictionaries to maintain words for each domain-specific concept.

From Figure 4 we can further observe that CTI-MTL achieves the best performance (0.8731 in micro-AUC) among all the comparison methods in inferring concept

transitions in medical queries. The CTI-MTL model has a nearly 2.5% improvement on micro-AUC when compared with coCTI and a nearly 7.5% improvement with CTI. This demonstrates that the mutual transfer loss which penalizes conflicts between the inferred concepts and inferred concept transitions can improve the inference quality.

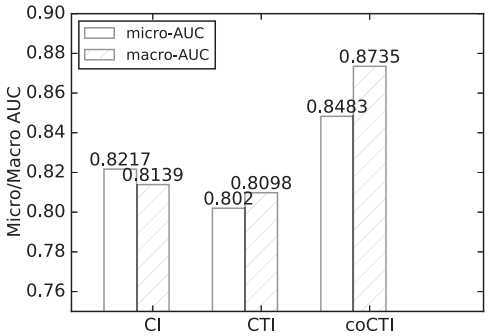

Figure 5. Micro/Macro-AUC scores for collective inference (coCTI) VS. concept inference(CI) and concept transition inference (CTI) separately.

Concept Transition	LR	NNID-JM	CTI	coCTI	coCTI-MTL
Symptom→Diet	0.6544 (5)	0.7755 (4)	0.7669 (3)	0.7959 (2)	0.8495 (1)
Symptom→Medicine	0.7022 (5)	0.7893 (4)	0.8242 (3)	0.8571 (2)	0.8624 (1)
Symptom→Cause	0.7600 (5)	0.8549 (4)	0.8786 (3)	0.8911 (1)	0.8880 (2)
Disease→Diet	0.7818 (5)	0.8670 (4)	0.8681 (3)	0.9059 (2)	0.9458 (1)
Disease→Treatment	0.7181 (5)	0.7787 (3)	0.7482 (4)	0.8456 (2)	0.8836 (1)
Disease→Examine	0.6397 (5)	0.6707 (4)	0.7838 (3)	0.8221 (2)	0.8480 (1)
Disease→Medicine	0.7623 (5)	0.8726 (4)	0.8749 (3)	0.8873 (2)	0.9015 (1)
Surgery→Recover	0.8117 (5)	0.9126 (3)	0.9012 (4)	0.9239 (2)	0.9396 (1)
Surgery→Sequela	0.7385 (5)	0.8031 (4)	0.8214 (3)	0.8417 (2)	0.8972 (1)
Surgery→Syndrome	0.7896 (5)	0.7994 (4)	0.8634 (2)	0.8619 (3)	0.9172 (1)
Surgery→Risk	0.6613 (5)	0.8063 (4)	0.8688 (3)	0.8715 (2)	0.9099 (1)
Medicine→Symptom	0.6861 (5)	0.8275 (3)	0.7553 (4)	0.8294 (2)	0.8598 (1)
Medicine→Side Effect	0.6652 (5)	0.8162 (3)	0.7771 (4)	0.8135 (2)	0.8814 (1)
Medicine→Disease	0.6806 (4)	0.6514 (5)	0.8081 (3)	0.8126 (2)	0.8678 (1)
Medicine→Instruction	0.7090 (5)	0.7761 (3)	0.7603 (4)	0.8170 (2)	0.8820 (1)
Examine→Fee	0.7576 (5)	0.9049 (3)	0.8981 (4)	0.9425 (2)	0.9482 (1)
Examine→Diagnosis	0.6832 (5)	0.7956 (3)	0.7445 (4)	0.8383 (2)	0.8822 (1)
Symptom→Treatment	0.6817 (5)	0.7640 (3)	0.7313 (4)	0.8130 (2)	0.8531 (1)
Symptom→Department	0.5978 (5)	0.6460 (3)	0.6013 (4)	0.6738 (2)	0.8080 (1)
Disease→Cause	0.7306 (5)	0.8206 (4)	0.8515 (3)	0.8608 (2)	0.8634 (1)
Disease→Symptom	0.6936 (4)	0.7552 (3)	0.6845 (5)	0.7554 (2)	0.8372 (1)
Disease→Department	0.6931 (5)	0.7387 (4)	0.7431 (3)	0.7652 (2)	0.8290 (1)
Disease→Surgery	0.7801 (5)	0.8795 (4)	0.9029 (3)	0.9236 (2)	0.9380 (1)

TABLE 2. FINE-GRAINED AUC SCORES FOR CONCEPT TRANSITION INFERENCE FOR EACH CONCEPT TRANSITION (EACH EDGE IN THE CONCEPT GRAPH).

Figure 5 shows the effectiveness of the co-inference procedure by comparing the performance of CTI with coCTI. The CI infers concept mentions so we can't simply compare its performance with CTI/coCTI where concept transitions are inferred. However, for CTI and coCTI, the improved performance on both micro-AUC and macro-AUC validate the effectiveness of inferring concept transitions and concept mentions collectively than inferred separately. The coCTI model can be considered as a multi-task learning model where the question representation is learned jointly and shared between two inference tasks.

Furthermore, the fine-grained AUC scores on all concept transitions without micro/macro-averaging are shown in Table 2. A general observation we can draw from the results is that the coCTI-MTL model is able to outperform other baselines in almost all types of concept transitions.

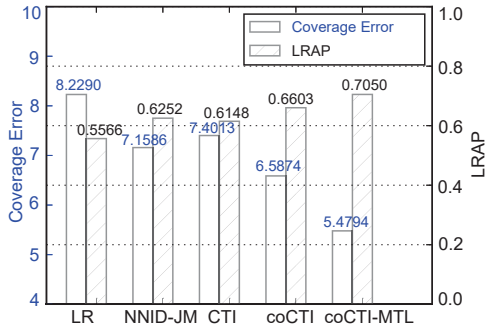

Figure 6. Coverage Loss and Label Ranking Average Precision (LRAP).

Figure 6 shows the coverage loss and LRAP over proposed methods and other baselines, where the coCTO-MTL model is able to achieve the lowest coverage error and the highest label ranking average precision score.

4. Related Works

4.1. Medical Query Analysis

As a growing number of people are posting medical related questions or searching with medical text queries online, researchers have been focusing on new problems and applications based on medical queries or search queries that users generated. [25] analyzes the conceptual relationship in medical records for a better medical search. [26] studies the circumlocution problem in diagnostic medical queries, where users are not able to express their ideas effectively. [4] tries to model user intentions as a classification task for medical text queries. [27] proposes a technique to detect whether users express patient experiences in their medical text queries. [28] introduces medical knowledge discovery from online question-answering corpus. In [29], authors introduce a neural network model to understand users healthcare related questions and try to generate answers appropriately. Being able to infer medical concept transitions from noisy, user-generated healthcare questions may further facilitate various medical applications such as healthcare question-answering, medical dialog systems or recommendation. For example, once we extracted the concept transition $Symptom \rightarrow Medicine$ from a question *Any medication is recommended to help me fall asleep easier?*, we may follow up by recommending the user to the nearest pharmacy for further medical consultations on corresponding OTC medicines on Insomnia.

4.2. Text Classification

Recently, lots of neural network models are developed for classifying natural language texts into different categories [30], [31]. Those methods achieve decent performance on general text classification tasks. The proposed concept transition problem can be cast as a multi-class multi-label classification problem. Unlike traditional text classification tasks like news classification where the existence of some topic words may easily dominate the label for a news title, users tend to mention multiple medical

concepts in a single medical text query. It is crucial to extract user medical concept transitions among multiple medical concepts, besides just concept mentions individually.

Also, the aforementioned methods consider the textual information only. With a graph-based formation in this paper, our model is able to seamlessly incorporates an existing concept graph with text information. Moreover, we propose to predict concept mentions as nodes and transitions as links on an abstract level collectively, while most existing works have been focusing on predicting links among concrete entities, e.g. among users in social networks [32], or predicting links among entities on a knowledge graph [33], [34].

5. Conclusions

People nowadays are posting or searching with medical text queries extensively on the world wide web. Various medical information needs are expressed diversely in users medical text queries. In this work, we bring semantic structures to user intention detection in real-world online medical queries by mapping diversely expressed medical queries to a concept graph where each node on a concept graph represents a concept mention and concept transitions are represented as directed edges. A novel neural network structure based on multi-task learning is introduced to extract concept mentions as well as medical concept transitions that users encoded in online healthcare questions collectively. Evaluation results on real-world medical questions address the effectiveness of the proposed model.

6. Acknowledgments

This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432, and NSFC 61672313.

References

[1] J. Campbell, G. Dussault, J. Buchan, F. Pozo-Martin, M. Guerra Arias, C. Leone, A. Siyam, and G. Cometto, "A universal truth: no health without a workforce," *Geneva: World Health Organization*, 2013.

[2] S. Fox and M. Duggan, "One in three american adults have gone online to figure out a medical condition," *Pew Internet & American Life Project*, 2013.

[3] M. Zhang and C. C. Yang, "Classification of online health discussions with text and health features sets," in *W3PHI*, 2014.

[4] C. Zhang, W. Fan, N. Du, and P. S. Yu, "Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach," in *WWW*, 2016.

[5] C. Zhang, N. Du, W. Fan, Y. Li, C.-T. Lu, and P. S. Yu, "Bringing semantic structures to user intent detection in online medical queries," *arXiv preprint arXiv:1710.08015*, 2017.

[6] A. De and S. K. Kopparapu, "A rule-based short query intent identification system," in *ICSIP*. IEEE, 2010, pp. 212–216.

[7] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu, "A study of medical and health queries to web search engines," *Health Information & Libraries Journal*, vol. 21, no. 1, pp. 44–51, 2004.

[8] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Use of sentiment analysis for capturing patient experience from free-text comments posted online," *JMIR*, vol. 15, no. 11, 2013.

[9] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, "Can I hear you? sentiment analysis on medical forums." in *IJCNLP*, 2013.

[10] J. Legrand and R. Collobert, "Joint rnn-based greedy parsing and word composition," in *ICLR*, 2015.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[14] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

[18] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *AAAI*, 2016.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.

[20] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[21] J. chinese word segmentation package. https://github.com/fxsjy/jieba.

[22] H.-P. Zhang, H.-K. Yu, D.-Y. Xiong, and Q. Liu, "Hhmm-based chinese lexical analyzer ictclas," in *SIGHAN*. Association for Computational Linguistics, 2003, pp. 184–187.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *AISTATS*, vol. 9, 2010, pp. 249–256.

[25] N. Limsopatham, C. Macdonald, and I. Ounis, "Inferring conceptual relationships to improve medical records search," in *OAIR*, 2013.

[26] I. Stanton, S. Ieong, and N. Mishra, "Circumlocution in diagnostic medical queries," in *SIGIR*. ACM, 2014.

[27] Y. Liu, Y. Chen, J. Tang, and H. Liu, "Context-aware experience extraction from online health forums," in *ICHI*.

[28] Y. Li, C. Liu, N. Du, W. Fan, Q. Li, J. Gao, C. Zhang, and H. Wu, "Extracting medical knowledge from crowdsourced question answering website," *IEEE Transactions on Big Data*, 2016.

[29] C. Liu, H. Sun, N. Du, S. Tan, H. Fei, W. Fan, T. Yang, H. Wu, Y. Li, and C. Zhang, "An augmented lstm framework to construct medical self-diagnosis android," in *ICDM*, 2016.

[30] E. Grefenstette and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," *ACL*, 2014.

[31] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," *AAAI*, pp. 2267–2273, 2015.

[32] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *JASIST*, vol. 58, no. 7, pp. 1019–1031, 2007.

[33] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2016.

[34] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *NIPS*, 2013, pp. 2787–2795.