Use of "Internal Knowledge": Biomedical Literature Search Liberated From External Resources

Tianwen Jiang[‡*], Ning Zhang[‡*], Ming Liu[‡], Meng Jiang[†], Ting Liu[‡], Bing Qin[‡]
[‡]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[†]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
{twjiang, nzhang, mliu, tliu, bqin}@ir.hit.edu.cn, mjiang2@nd.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313110

***Abstract*—Knowledge plays an essential role in biomedical literature search (BLS) systems, filling the semantic gap between queries and documents. Knowledge bases, constructed by human experts or machine learning methods, are generally regarded as the main sources serving *external* knowledge. However, a good knowledge base must balances its particularity and generalization, resulting limited knowledge coverage and utilization to BLS systems. Considering massive documents in a BLS system, and recently developing Open IE techniques by which we can automatically extract structured knowledge from documents, how about harnessing distilled *internal* knowledge rather than external knowledge to conduct BLS systems? Internal knowledge, providing tailored particular knowledge to BLS systems, is supposed to lead better knowledge utilization and much more competitive performance on literature search. In this paper, we design an novel internal knowledge driven BLS system upon a Multi-layered Encoders incorporating Multi-layered internal Knowledge graph, called MEMK. MEMK harnesses distilled internal structural knowledge, empowering interactive representations learning of query and documents. The experiments show that MEMK outperforms strong baselines on a public benchmark, and internal knowledge based query expansion can further improve the performance to a new state of the art.**

I. Introduction

Biomedical literature search (BLS) systems are targeted at providing scientists with quick access to massive biomedical literature database for useful information [1]–[3], as biomedical research develops rapidly in biomedical domain. Biomedical domain knowledge plays an essential role in BLS, by reducing the semantic gap between query and documents and significantly improving the performance of a BLS systems [4], [5]. Biomedical knowledge bases (BioKBs) (e.g., MeSH, GO, and Uniprot), manually or automatically pre-constructed, are widely utilized to serve domain ***external*** knowledge.

However, ***external*** knowledge, delivered from BioKBs, gets limitations in two aspects for knowledge use: ***completion*** and ***advancement***. First, besides time-consuming of human-experts constructed BioKBs, both of manually and automatically built BioKBs suffer from being incomplete [6], [7]. Essential knowledge required by BLS systems to fill semantic gap between query and documents is often missing when looking up from BioKBs. Second, most present BioKBs are constructed to provide general and classical common biomedical knowledge, which usually are not staying up to date, resulting limited advancement of BioKBs. Current BLS systems (e.g., PubMed [8]) can only benefit from lexical knowledge (e.g., hypernyms and synonyms of items) from most of BioKBs.

We consider harnessing another knowledge source rather than external BioKBs: query text and documents text in BLS system, inducing ***internal knowledge***. Internal knowledge can be automatically distilled from query and documents text, thanks to recent development of Open IE techniques [9] in biomedical domain. Internal knowledge provides promising ***completion*** and ***advancement*** of knowledge to BLS system. First, massive domain related literatures in BLS system serve sufficient knowledge to fill semantic gap between query text and candidate documents text, and internal knowledge is exactly induced from these literatures. Second, advancement of internal knowledge is determined by content of BLS systems literatures, so as we keep BLS system up to the date, advancement of knowledge use is promised naturally. Internal knowledge presents an attractive opportunity for the BLS system's demand for knowledge and improvement of performance, both in quality and quantity.

Present work. We propose to use internal knowledge for BLS in two main aspects, namely, query expansion and representation learning over queries and document collection. In this work, the internal knowledge is extracted by a open-source biomedical OpenIE system, called MIMO [9], which has proved great efficacy in biomedical retrieval tasks [10]. The internal knowledge is structural in three layers: concept layer, tuple layer, and top statement layer. First, for a given query in query expansion, a set of factual related terms (concepts) will be retrieved from massive fact tuples of internal knowledge over document collection, and we append them to expand original query. Second, compared to feature engineering, end-to-end representation learning has achieved extraordinary results on many natural language processing tasks. We propose a new joint representation learning method of text and structured internal knowledge, in four different information granularity: from plain text, concepts, and tuples to statements, modeling the semantic relevance of queries and their candidate documents.

We propose an external resources free neural BLS system: A **M**ulti-layered **E**ncoders that hierarchically learns the representation of text content and **M**ulti-layered internal **K**nowledge, equipped with internal knowledge based query expansion, called **MEMK**. MEMK has two core modules, corresponding

*Equal contribution.

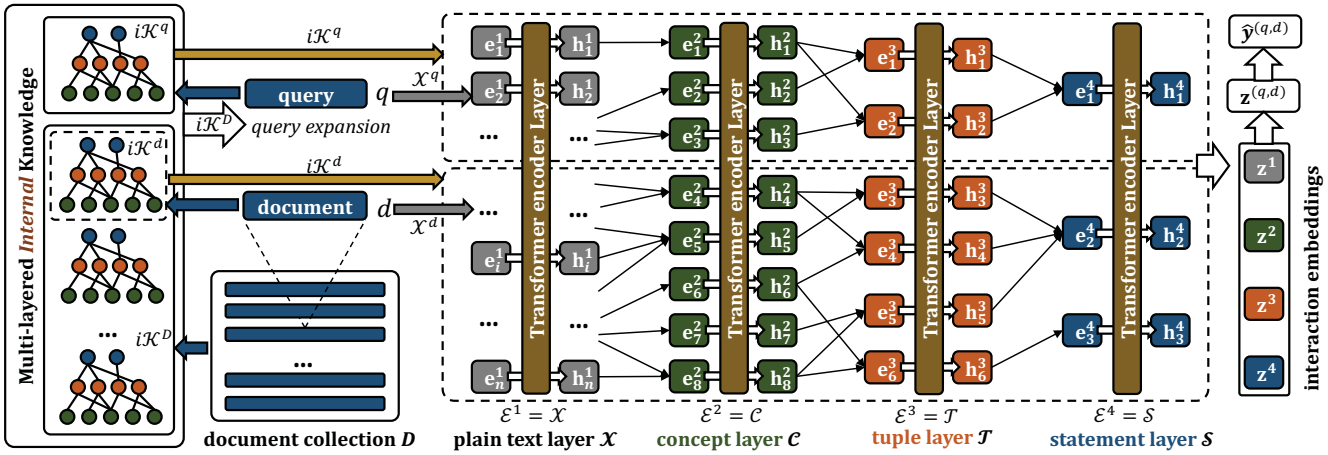

Fig. 1: The proposed MEMK method utilizes multiple Transformer encoder layers to model plain text content and a multi-layered internal knowledge graph. Four stacked Transformer encoder layers are for plain text layer, concept layer, tuple layer and statement layer, respectively.

to two ways of internal knowledge use: query expansion module, representation learning module. For query expansion module, MEMK matches the internal knowledge of query and the document collection to obtain the factual related concepts to expand the original queries. In the representation learning module, MEMK uses multiple stacked Transformer encoder layers to model four types of information: plain text content, the concept layer, tuple layer and statement layer of internal knowledge. To our best knowledge, MEMK is the first to incorporate structured internal knowledge into neural network models and applied it to biomedical literature search task, requiring no external knowledge resources. In experiments, MEMKs outperform strong baselines on a public benchmark: MEMK without query expansion significantly improve the previous best performances of the benchmark on F-measure, MAP measure, relatively by 16.92%, 14.78%; with the help of the internal knowledge based query expansion, MEMK achieves a new state of the art performance, further improving the original MEMK relatively by 4.20%, 1.66% on F-measure, MAP measure.

II. Problem Statement

Definition 1 (Biomedical Literature Search. Given a biomedical text query q, and a document collection of biomedical domain $D = \{d_1, d_2, \cdots, d_N\}$ with N documents. A biomedical literature search (BLS) system is supposed to return a ranked search results of biomedical documents list $[d'_1, d'_2, \cdots, d'_N]$ for q, where the higher ranks represent the more likely to be relevant.

Definition 2 (Multi-layered Internal Knowledge). We apply Open IE system to the text content of text query q and document collection D, to obtain a set of (subject, relation, object)-tuples, as *internal* knowledge. Formally, we use a three-layered graph to represent the internal knowledge following previous work [11], [12], denoted as $i\mathcal{K}$:

$$i\mathcal{K} = \{\mathcal{C}, \mathcal{T}, \mathcal{S}, E_{\mathcal{C},\mathcal{T}}, E_{\mathcal{T},\mathcal{S}}\} \quad (1)$$

where $\mathcal{C}$ is the first layer of concepts, as subjects or objects of the tuples in the second layer. $\mathcal{T}$ is the second layer of tuples. Each tuple is represented as the node of relation name, connecting to its subject and object in the first layer, specified by the edges in edge set $E_{\mathcal{C},\mathcal{T}}$. $\mathcal{S}$ is the third layer of statements. Each statement node connect to a set of tuples in the second layer, specified by the edges from the edge set $E_{\mathcal{T},\mathcal{S}}$. For a specific query q, we denote its internal knowledge as $i\mathcal{K}^q$; for a document d_i, we denote its internal knowledge as $i\mathcal{K}^{d_i}$. Document collection D gets a internal knowledge set as $i\mathcal{K}^D = \{i\mathcal{K}^{d_1}, i\mathcal{K}^{d_2}, \cdots, i\mathcal{K}^{d_N}\}$.

Problem (Neural BLS using Multi-layered Internal Knowledge). Given a biomedical text query q and a document collection of biomedical domain $D = \{d_1, d_2, \cdots, d_N\}$, and their multi-layered internal knowledge $i\mathcal{K}^q$ and $\{i\mathcal{K}^{d_1}, i\mathcal{K}^{d_2}, \cdots, i\mathcal{K}^{d_N}\}$, learn an interaction based neural networks that can return a ranked search results of biomedical documents list $[d'_1, d'_2, \cdots, d'_N]$ for q, where the higher-ranked documents represent the more likely the target documents. Here $i\mathcal{K}$ is multi-layered graph of concept set $\mathcal{C}$, tuple set $\mathcal{T}$ and statement set $\mathcal{S}$: $i\mathcal{K} = \{\mathcal{C}, \mathcal{T}, \mathcal{S}, E_{\mathcal{C},\mathcal{T}}, E_{\mathcal{T},\mathcal{S}}\}$.

III. The Proposed Approach

A. MEMK

In this section, we introduce MEMK, an external knowledge free BLS model: **M**ulti-layered **E**ncoders incorporating **M**ulti-layered internal **K**nowledge graph. MEMK uses an interaction based neural approach, taking the concatenated plain text in query q and document d as input. Each encoder module of MEMK is the stacked self-attention and point-wise fully connected feed-forward networks [13]. MEMK gets four layers of encoders, with respect to plain text and three-layered internal knowledge. The input embeddings of each encoder layer are depended on the output embeddings of the previous encoder layer and connections to it. The overall framework is visualized in Figure 1.

Transformer encoder layer. We concatenate the plain query text $\mathcal{X}^q$ with the plain document text $\mathcal{X}^d$ as the input: the first encoder layer takes the concatenated plain text as input, so as the remaining encoder layer taking the corresponding type of concatenated information of the query q and the document d. We use notation $\mathcal{E}$ to represent the input sequence of the current encoder layer, where $\mathcal{E} \equiv [\texttt{CLS}; \mathcal{E}^q; \texttt{SEP}; \mathcal{E}^d]$. $\mathcal{E}^q \in \{\mathcal{X}^q, \mathcal{C}^q, \mathcal{T}^q, \mathcal{S}^q\}$ is the query q part of the concatenation, and $\mathcal{C}^q, \mathcal{T}^q, \mathcal{S}^q \in i\mathcal{K}^q$; $\mathcal{E}^d \in \{\mathcal{X}^d, \mathcal{C}^d, \mathcal{T}^d, \mathcal{S}^d\}$ is the document d part of the concatenation, and $\mathcal{C}^d, \mathcal{T}^d, \mathcal{S}^d \in i\mathcal{K}^d$. Here CLS, SEP are the beginning and separating sign.

$\mathcal{E}^q$ and $\mathcal{E}^d$ are both sequences: $\mathcal{E}^q \equiv [e_1^q, e_2^q, \cdots, e_n^q]$, and $\mathcal{E}^d \equiv [e_1^d, e_2^d, \cdots, e_m^d]$. e_*^* represent the element of the input sequence: text token, concept, tuple, or statement node. We write $\mathcal{E}$ as a sequence $[e_1, e_2, \cdots, e_{(n+m+2)}]$, taking the count of CLS and SEP. After going through the embedding layer or aggregating from outputs in the previous encoder layer, we get the continuous representation of $\mathcal{E}$ as a sequence of vectors, denoted as $\mathbf{E}$, where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{(n+m+2)}]$ and $\mathbf{e}_*$ is the vector of e_*. The Transformer encoder layer encodes the input embedding $\mathbf{E}$ as follows:

$$\widetilde{\mathbf{H}}^l = \mathrm{LN}(\mathbf{H}^{l-1} + \mathrm{MHAtt}(\mathbf{H}^{l-1})) \tag{2}$$

$$\mathbf{H}^l = \mathrm{LN}(\widetilde{\mathbf{H}}^l + \mathrm{PwFFN}(\widetilde{\mathbf{H}}^l)) \tag{3}$$

where LN is the layer normalization operation; MHAtt is the multi-head attention operation, and PwFFN is the point-wise feed-forward network. The superscript $l \in \{0, 1, \cdots, L\}$ indicates the depth of the stacked layer: $\mathbf{H}^0 \equiv \mathbf{E}$, and $\mathbf{H}^L$ is the final output of the Transformer encoder. $\mathbf{H}^L = [\mathbf{h}_1^L, \mathbf{h}_2^L, \cdots, \mathbf{h}_{(n+m+2)}^L]$, where $\mathbf{h}_i^L$ is the output embedding for the $i-$th information element e_i, i.e., token, concept, tuple, or statement node. We use $\mathbf{H}$ to represent $\mathbf{H}^L$ for short, and use $\mathbf{h}_i$ to represent $\mathbf{h}_i^L$ for short.

The final interaction embedding $\mathbf{z}$, to estimate the relevance of q and d in the current layer, is obtained by applying a pooling layer to the matrix $\mathbf{H}$:

$$\mathbf{z} = \mathrm{Pooing}(\mathbf{H}) \tag{4}$$

where we pool $\mathbf{H}$ by simply taking the hidden state corresponding to the first token, following a FFNN layer.

Progressive encoding from text to knowledge. In each encoder layer of MEMK, except the first encoder layer where input embeddings $\mathbf{E}$ are produced by embedding layer, input embeddings $\mathbf{E}$ is aggravated from the output embeddings $\mathbf{H}$ from previous encoder layer(s) guided by the layer connections of the internal knowledge graph $i\mathcal{K} = i\mathcal{K}^q \cup i\mathcal{K}^d$. We use $\mathcal{E}^k$ to represent the raw input of the $k-$th encoder layer, where $k \in \{1, 2, \cdots, K\}$. And the continuous representation of $\mathcal{E}^k$ is $\mathbf{E}^k$, presenting the input embedding of the $k-$th encoder layer, and $\mathbf{H}^k$ presents the output embedding of the $k-$th encoder layer. $\mathbf{E}^1$ is directly obtained from embedding layer, and for remaining encoder layers where $k \in \{2, 3, \cdots, K\}$, we get the $\mathbf{E}^k$ as follows:

$$\mathbf{E}^k = \mathrm{iKAggregation}(\{\mathbf{H}^1, \mathbf{H}^2, \cdots, \mathbf{H}^{k-1}\}, i\mathcal{K}, \mathcal{X}, \mathcal{E}^k) \tag{5}$$

where $i\mathcal{K} = i\mathcal{K}^q \cup i\mathcal{K}^d$, and $\mathcal{X} = \mathcal{X}^q \cup \mathcal{X}^d$. For the raw input of the current encoder layer, to get its input embeddings, iKAggregation aggregates the targeted previous output embeddings according to two aspects of links information: inter-layer links (i.e., $E_{\mathcal{C},\mathcal{T}}$ and $E_{\mathcal{T},\mathcal{S}}$) of internal knowledge, the grounding links (i.e., $E_{\mathcal{X},\mathcal{C}}$ and $E_{\mathcal{X},\mathcal{T}}$) from concepts/relations to the tokenized texts $\mathcal{X}$.

Global interaction based relevance. Through the progressive encoding over each encoder layers, MEMK gets the multiple interaction embeddings from K encoders: $\mathbf{Z} = [\mathbf{z}^1, \mathbf{z}^2, \cdots, \mathbf{z}^K]$. These interaction embeddings transmit different while complementary information to estimate the relevance of the query q and the document d. The final global interaction embedding is also obtained first using a stacked self-attention network and point-wise fully connected feed-forward network, as defined in Equation 2 and Equation 3, where $\mathbf{H}^0 \equiv \mathbf{Z}$. Then MEMK uses a pooling layer (as defined in Equation 4) to get the final interaction embedding, denoted as $\mathbf{z}^{(q,d)}$. The relevance score $\hat{y}^{(q,d)}$ from 0 to 1 for (q, d) is defined as:

$$\hat{y}^{(q,d)} = \mathrm{Sigmoid}(\mathbf{z}^{(q,d)}) \tag{6}$$

Ranked search results returning. For each $d_i \in D$, where D is the biomedical document collection $[d_1, d_2, \cdots, d_N]$, MEMK calculates the relevance $\hat{y}^{(q,d_i)}$ for the given query q. This way MEMK returns the relevance score list: $[\hat{y}^{(q,d_1)}, \hat{y}^{(q,d_2)}, \cdots, \hat{y}^{(q,d_N)}]$ for D. The search results are then sorted in descending order of score in the returned relevance score list: $[d'_1, d'_2, \cdots, d'_N]$.

B. Internal Knowledge for Query Expansion

Considering that document collection D is relatively static in a short time, the internal knowledge of these documents in D can be pre-constructed, while the query internal knowledge is realtime constructed. A large-scale internal knowledge graph $i\mathcal{K}^*$, by merging the internal knowledge $i\mathcal{K}^D = \{i\mathcal{K}^{d_1}, i\mathcal{K}^{d_2}, \cdots, i\mathcal{K}^{d_N}\}$ over D: $i\mathcal{K}^* = \cup_{i=1}^N i\mathcal{K}^{d_i}$, and $i\mathcal{K}^* = \{\mathcal{C}^*, \mathcal{T}^*, \mathcal{S}^*, E^*_{\mathcal{C},\mathcal{T}}, E^*_{\mathcal{T},\mathcal{S}}\}$.

Given concept $c_i \in \mathcal{C}^*$ could establish contact with another concept $c_j \in \mathcal{C}^*$ in $i\mathcal{K}^*$, via a relation $r \in \mathcal{T}^*$. Concepts c_i and concept c_j are tuple-based co-occurrence. For the given query q and its hierarchical internal knowledge $i\mathcal{K}^q$, we get the concept set $\mathcal{C}^q$ from $\mathcal{K}^q$. Each $c \in \mathcal{C}^q$ will be matched to the large-scale knoweldge graph $i\mathcal{K}^*$, to get its concepts of tuple-based co-occurrence. The concepts of high-frequency co-occurrence are left as the candidates for query expansion.

C. Training for MEMK

For any input query-document pair (q, d) and its relevance label $y^{(q,d)}$, where 1 for positive query-document pair and 0 for negative query-document pair. We use cross-entropy loss function to training the MEMK model, where the loss is defined as:

$$\mathcal{L} = y^{(q,d)} \log(\hat{y}^{(q,d)}) + (1 - y^{(q,d)}) \log(1 - \hat{y}^{(q,d)}) \tag{7}$$

where $\hat{y}^{(q,d)}$ is the predicted label by MEMK.

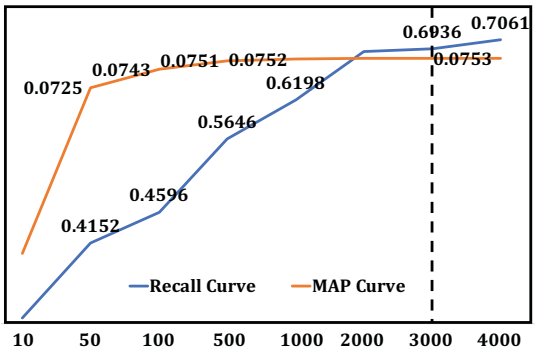

Fig. 2: Recall and MAP measures for PubMed retrieved documents on BioASQ Task 5b.

IV. Experiments

A. Experiments Setups

Datasets and Evaluation. According to BioASQ official published results paper [14], (1) we use the third test set batch of Task 5b (Task 5b-03) as test set, and use the fifth test set batch (Task 5b-05) as development set. (2) For the training sets, we use the training sets from 2013 to 2017 to train the models following the previous work in Task 5b. There are four types of biomedical questions: yes/no, factoid, list and summary questions. Document collection for BLS are PubMed Annual Baseline Repository for 2017. Primary evaluation metrics are mean precision (Prec.), mean recall (Recall), f-measure (F), MAP and GMAP as previous work [4], [14].

Competitive methods. We select six systems submitted to the BioASQ Task 5b as baselines, which have the six best MAP measures according to the published report paper [14]: **USTB** [4], **fdu** [14], **UNCC**, **Olelo**, **testtext**, and **HPI-S**. We also investigate other four related and recent baselines: (1) **PubMed** [8], as one of the most popular BLS system, is taken into the comparison. (2) **GloVe** [15], a pre-trained word embedding based method. (3) **CTGA** [10] uses graph-based method for BLS, using rule-based method to match the internal knowledge graph of queries and documents. (4)**BERT** [13], as a renown pre-training language model, demonstrates great improvement on information retrieval tasks [16]–[18].

B. Pre-processing

Candidates Retrieval. We take a two-steps document search: retrieval and re-ranking. In retrieval stage, we utilize PubMed [8] to obtain candidate documents from the document collection. We investigate the target documents distribution by measuring the top-K documents returned by PubMed, as shown in Figure 2, where the value of the x-axis represents using the top-K retrieved documents as the candidates. As the value of K increases, the mean recall value gradually increases as expected (around 70% is achieved when K=4000), but the MAP value tends to stabilize. We set K to 3000 to get the candidate documents by PubMed search engine. After removing the documents published latter than 2017 when BioASQ Task 5b held, each question query has about 2721 candidate documents, averagely. Our implemented methods including GloVe, BERT and MEMKs are all built on the PubMed providing candidate documents.

Methods	Mean Prec.	Mean Recall	Mean F	MAP	GMAP
HPI-S [14]	0.0823	0.2152	0.0997	0.0464	0.0005
Olelo [14]	0.1327	0.2444	0.1481	0.0658	0.0005
UNCC [14]	0.2317	0.3340	0.2322	0.0825	0.0009
fdu [14], [19]	0.1615	0.4475	0.2120	0.1021	0.0049
testtext [14]	0.1610	0.4690	0.2087	0.1138	0.0048
USTB [4], [14]	0.1620	0.4803	0.2111	0.1157	0.0050
PubMed [8]	0.0997	0.3082	0.1272	0.0567	0.0010
GloVe [15]	0.0844	0.3215	0.1069	0.0679	0.0009
CTGA [10]	0.1107	0.3350	0.1424	0.0878	0.0029
BERT [16]–[18]	**0.2635**	0.3278	0.2461	0.0940	0.0016
MEMK	0.2300	0.5160	0.2715	0.1328	0.0080
MEMK(Ex)	0.2430	**0.5322**	**0.2829**	**0.1350**	**0.0081**

TABLE I: The proposed MEMK methods, i.e., MEMK and MEMK(Ex), outperform existing methods in terms of Mean F measure, MAP measure, and GMAP measure on the document retrieval of BioASQ Task 5b. The results demonstrate the effectiveness of MEMKs and internal knowledge use.

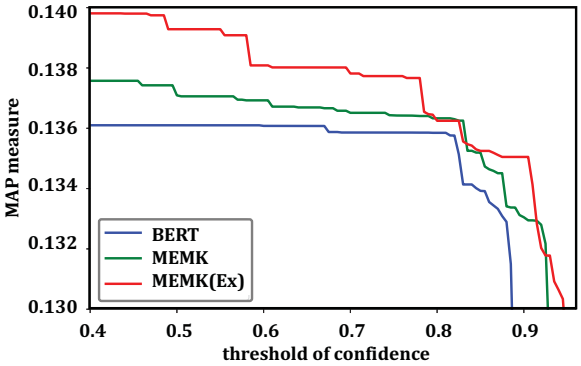

Fig. 3: Visualizing performance comparison of BERT, MEMK, and MEMK(Ex) on MAP measure of BioASQ Task5b document retrieval. (Best viewed in color.)

Internal knowledge obtaining. We utilize the open source of MIMO tool [9] to acquire the multi-layered internal knowledge for queries and candidate documents. We reserve both factual tuples and condition tuples, treating them equally in MEMK, as condition tuples are as important as factual tuples for matching the target document in biomedical domain. For example, in the query "Which cells express CIDEC protein in humans?", the condition of "in humans" is crucial to find the exact answerable documents. Finally, each question query gets 1.01 statements, 4.27 tuples, and 5.16 concepts in its multi-layered internal knowledge, averagely; each document gets 9.04 statements, 41.31 tuples, and 66.53 concepts in its multi-layered internal knowledge, averagely.

C. Results

Effectiveness of MEMKs and internal knowledge use. The MEMK models, including MEMK and MEMK with query expansion (MEMK(Ex)), outperform all the competitive methods, shown as Table I: MEMK incorporating internal knowledge significantly improve the previous best performances of

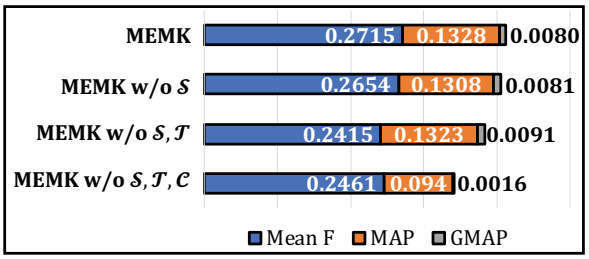

Fig. 4: Visualizing performance ablation study on MEMK.

the benchmark on F-measure, MAP measure, relatively by 16.92%, 14.78%; with the help of the internal knowledge based query expansion, MEMK achieves a new state of the art performance, further improving the original MEMK relatively by 4.20%, 1.66% on F-measure, MAP measure. Table I places the methods by MAP measure value in each blocks: the competitive methods of BioASQ, our implemented baselines, and our proposed methods. We have four other observations on Table I. *First*, USTB achieves the best MAP and GMAP performance, where external knowledge play an essential role. *Second*, UNCC achieves the best Precision and F measures among the competitors on the document retrieval of BioASQ Task 5b. *Third*, BERT, similar to UNCC, utilizes the internal latent semantic information for BLS task. *Fourth*, PubMed achieves not well performance on the document retrieval task of BioASQ, because PubMed was designed for efficient paper search by a few keywords such as author name and paper title, not for question queries search.

Effectiveness of the multi-layered internal knowledge. The internal knowledge MEMK used is multi-layered, and we design a multi-layered encoders in MEMK to model the multi-layered internal knowledge. Here we investigate the effects of different layers in the multi-layered internal knowledge: concept layer, tuple layer, and statement node layer. For the original MEMK model, we gradually eliminate statement node layer, tuple layer, and concept layer, concerning the performance after each elimination of the layers. Figure 4 shows these results as ablation study, where we can find three observations. (1) *First*, MEMK that utilizes all the knowledge layer, i.e., MEMK, has the best F measure and MAP measure, leading relative 2.30% ahead on F measure of the best model (MEMK w/o $\mathcal{S}$) that only utilizes partial knowledge layers. (2) *Second*, the ablation of statement node layer causes the least performance loss. Statement node layer provides a structure knowledge about the tuple distribution over statements, for most cases such structure knowledge is not as essential as semantic knowledge about concepts and tuples. (3) *Third*, knowledge layers of concepts and tuples play an essential role in MEMK: relative 9.00% drop on F measure brought by the ablation of tuple layer; relative 28.13% drop on MAP measure brought by the ablation of concept layer. The MEMK has no internal knowledge, only utilizing plain text, is exactly the BERT model in the experiments, leading a poor MAP measure.

Effectiveness on different types of questions. There are four types of question quires in BioASQ: yes/no, factoid, list and summary. We are interested in the actually performance on each type of quires of the methods. We conduct such experiments on BERT, MEMK, and MEMK(Ex) methods, as we cannot obtain the detailed results of the competitors on BioASQ Task 5b and BERT has the best F measure among the baselines. Table II present the results, where we have the following observations. (1) *First*, both MEMK and MEMK(Ex) outperform the BERT method by F, MAP, and GAMP measure on most of question quires. Especially, for the Recall measure, the MEMKs significantly lead 4.24% to 28.75% ahead of BERT. Internal knowledge has a great positive effects on recalling more target documents by different types of knowledge. (2) *Second*, MEMK(Ex) outperforms MEMK by F, MAP, and GMAP measures on the three of the four types of question queries, except "yes/no" queries. It demonstrates that internal knowledge based query expansion has positive effects on the most of question queries. In "yes/no" queries, all the crucial phrases that effect literature search are present in queries, this way query expansion is not quite needed and may even hurt the Precision by recalling more documents.

Methods	Mean Prec.	Mean Recall	Mean F	MAP	GMAP
Type: list (15)					
BERT [16]–[18]	0.2582	0.2422	0.2280	0.0480	0.0020
MEMK	0.2600	0.2644	0.2367	0.0559	0.0022
MEMK(Ex)	**0.2606**	**0.3450**	**0.2399**	**0.0872**	**0.0050**
Type: factoid (26)					
BERT [16]–[18]	**0.2484**	0.3489	0.2469	0.0508	0.0016
MEMK	0.2320	0.3736	0.2427	0.0567	0.0019
MEMK(Ex)	0.2190	**0.3913**	**0.2569**	**0.0645**	**0.0019**
Type: summary (28)					
BERT [16]–[18]	0.2779	0.2965	0.2338	0.1928	0.0016
MEMK	0.3078	0.5483	0.3468	0.2495	0.0096
MEMK(Ex)	**0.3571**	**0.5840**	**0.3792**	**0.2494**	**0.0133**
Type: yes/no (31)					
BERT [16]–[18]	0.2657	0.3799	0.2652	0.0633	0.0015
MEMK	**0.282**8	0.6053	**0.3299**	**0.1112**	**0.0121**
MEMK(Ex)	0.2774	**0.6061**	0.3293	0.1083	0.0102

TABLE II: The proposed MEMK methods, i.e., MEMK and MEMK(Ex), outperform BERT method in terms of Mean F measure, MAP measure, and GMAP measure on all the four query types. (Higher score is better)

V. Related Work

Neural approaches to ad-hoc IR. Neural ad-hoc IR models [16], [17] that perform search in a latent space tend to be more robust towards the term mismatch issue compared to lexical term-based ad-hoc IR models. Some methods used unsupervised embeddings to generate suitable query expansion candidates from a global vocabulary and then perform retrieval based on the expanded query [20], [21]. Semi-supervised neural approach using limited training data has been widely applied, such as [22], [23]. Supervised Siamese networks based approach [24], [25] learns the embeddings for query and document independently. Interaction based neural approach [26], [27] learns interaction vectors over different

parts of query and document for an aggregated relevance estimation, instead of learning the explicit representations of query and document. Some BERT-based model [16], [17] takes attention as the interaction function to learn the interaction vector between inputs.

Biomedical literature search. Balancing the semantic gap between query and documents is the crucial key for BLS task, and thus biomedical domain background knowledge is desired [5], [28]. Oleynik *et al.* [3] used external knowledge, such as NCBI gene list, to expand the original query by its preferred term, synonyms and hypernyms etc. Zhou *et al.* [5] developed a specific knowledge graph as external knowledge, and used the KG to obtain related concepts for original query. USTB [4] systems combined different strategies to enrich query terms, using five ontologies or terminologies as external knowledge: MeSH, GO, Jochem, Uniprot and DO. In addition to use MetaMap to extract concepts from query and get their name variants, Agosti *et al.* [29] adapted a pseudo relevance feedback query expansion technique to get more relevant concepts for original query. Taken together, various methods all attempted to enrich original query with more relevant knowledge or information, by query expansion and pseudo relevance feedback techniques.

VI. Conclusion

In this work, we proposed the idea that internal knowledge hidden in queries and documents can already provide sufficient biomedical knowledge for BLS task, thus external free BLS methods are desired. We designed a **M**ulti-layered **E**ncoders incorporating **M**ulti-layered internal **K**nowledge graph, called MEMK. MEMK utilizes internal structural knowledge for the representation learning of query and documents for BLS task. Experiments on a public BLS dataset showed that MEMK outperformed existing methods and internal knowledge based query expansion can further improve the performance of MEMK as a new state of the art.

Acknowledgments

This work was supported in part by the Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Major Project (2018AA0101901), the National Key Research and Development Project (2018YFB1005103), the National Natural Science Foundation of China (61772156, 61976073).

References

[1] G. Tsatsaronis and G. e. Balikas, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, p. 138, 2015.

[2] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, and I. Kakadiaris, "Results of the sixth edition of the BioASQ challenge," in *BioASQ*, Nov. 2018, pp. 1–10.

[3] M. Oleynik, E. Faessler, A. M. Sasso, A. Kappattanavar, B. Bergner, H. F. Da Cruz, J.-P. Sachs, S. Datta, and E. Böttinger, "Hpi-dhc at trec 2018 precision medicine track," in *TREC*, 2018.

[4] Z.-X. Jin, B.-W. Zhang, F. Fang, L.-L. Zhang, and X.-C. Yin, "A multi-strategy query processing approach for biomedical question answering: USTB_PRIR at BioASQ 2017 task 5B," in *BioNLP 2017*. Vancouver, Canada,: Association for Computational Linguistics, Aug. 2017, pp. 373–380.

[5] X. Zhou, X. Chen, J. Song, G. Zhao, and J. Wu, "Team cat-garfield at trec 2018 precision medicine track," in *TREC*, 2018.

[6] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in neural information processing systems*, 2013, pp. 926–934.

[7] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou, "End-to-end structure-aware convolutional networks for knowledge base completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3060–3067.

[8] N. Fiorini, K. Canese, G. Starchenko, E. Kireev, W. Kim, V. Miller, M. Osipov, M. Kholodov, R. Ismagilov, S. Mohan *et al.*, "Best match: new relevance search for pubmed," *PLoS biology*, vol. 16, no. 8, 2018.

[9] T. Jiang, T. Zhao, B. Qin, T. Liu, N. Chawla, and M. Jiang, "Multi-input multi-output sequence labeling for joint extraction of fact and condition tuples from scientific text," in *EMNLP-IJCNLP*. Hong Kong, China: ACL, Nov. 2019.

[10] T. Jiang, Z. Zhang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, and M. Jiang, "Ctga: Graph-based biomedical literature search," in *BIBM*, 2019, pp. 395–400.

[11] T. Jiang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, and M. Jiang, "The role of" condition" a novel scientific knowledge graph representation and construction model," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1634–1642.

[12] T. Jiang, Q. Zeng, T. Zhao, B. Qin, T. Liu, N. Chawla, and M. Jiang, "Biomedical knowledge graphs construction from conditional statements," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2020.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, and I. Kakadiaris, "Results of the fifth edition of the bioasq challenge," in *BioNLP*, 2017, pp. 48–57.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[16] W. Yang, H. Zhang, and J. Lin, "Simple applications of bert for ad hoc document retrieval," *arXiv preprint arXiv:1903.10972*, 2019.

[17] Z. Akkalyoncu Yilmaz, S. Wang, W. Yang, H. Zhang, and J. Lin, "Applying BERT to document retrieval with birch," in *EMNLP-IJCNLP*, Nov. 2019, pp. 19–24.

[18] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.

[19] Y. Zhang, S. Peng, R. You, Z. Xie, B. Wang, and S. Zhu, "The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering," in *CEUR Workshop Proceedings*, vol. 1391, 2015.

[20] F. Diaz, B. Mitra, and N. Craswell, "Query expansion with locally-trained word embeddings," *arXiv preprint arXiv:1605.07891*, 2016.

[21] D. Roy, D. Paul, M. Mitra, and U. Garain, "Using word embeddings for automatic query expansion," *arXiv preprint arXiv:1606.07608*, 2016.

[22] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition," in *AAAI*, 2016.

[23] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in *CIKM*, 2016, pp. 55–64.

[24] B. Mitra, F. Diaz, and N. Craswell, "Learning to match using local and distributed representations of text for web search," in *WWW*, 2017, pp. 1291–1299.

[25] H. Zamani, B. Mitra, X. Song, N. Craswell, and S. Tiwary, "Neural ranking models with multiple document fields," in *WSDM*, 2018, pp. 700–708.

[26] K. Hui, A. Yates, K. Berberich, and G. De Melo, "Co-pacrr: A context-aware neural ir model for ad-hoc retrieval," in *WSDM*, 2018, pp. 279–287.

[27] Y. Fan, J. Guo, Y. Lan, J. Xu, C. Zhai, and X. Cheng, "Modeling diverse relevance patterns in ad-hoc retrieval," in *SIGIR*, 2018, pp. 375–384.

[28] Z. Zheng, C. Li, B. He, and J. Xu, "Ucas at trec-2018 precision medicine track," in *TREC*, 2018.

[29] M. Agosti, G. M. D. Nunzio, and S. Marchesin, "The university of padua ims research group at trec 2018 precision medicine track," in *TREC*, 2018.