Multi-Granularity Retrieval Model for Bridging Gaps between Biomedical Concepts and Entities: THUIR at TREC 2007 Genomics Track

Hongning Wang, Jiao Li, Shilin Ding, Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China

Abstract

General concepts are always used to describe query requirement (In the example “What tumor types are associated with Rb1 mutations?”, “Tumor types” is a general concept, and its entity in a relevant documents can be “brain tumor”). To bridge the gaps between concepts in user queries and entities in relevant documents, we proposed a multi-granularity retrieval model in TREC 2007 Genomics task. The model consists of three components: (1) Paragraph retrieval is employed to retrieve candidate paragraph initially; (2) Dictionary-based NER is utilized to recognize named entities of given types; (3) Passage ranking is used to rank retrieved candidate passages. Our proposed model achieve promising result (Passage MAP=0.1023, with NER bottleneck eliminated).

1. Introduction

The major challenges of TREC 2007 Genomics Track are understanding queries which include desired entity types/general concepts (such as genes, proteins and disease), and retrieving passages which provide relevant information to such queries. There are 162,259 documents in the corpus from about 49 genomics related journals, which are the same with the last year’s full-text document corpus [1]. We proposed a multi-granularity retrieval model to solve the above challenges based on our previous work [2]. The rest of this paper is organized as following: Section 2 describes our methods in details; our results and system performances are analysis in Section 3; and the advantages and disadvantages of our model are discussed at the end.

2. Methods

We proposed multi-granularity retrieval model to, 1) retrieve potential paragraphs; 2) identify named entities; 3) rank the retrieved passages in the following workflow (as shown in Figure 1).

Figure 1. System Workflow Chart
In our system workflow, (1) Paragraph Retrieval module is designed to return relevant paragraphs \( (P) \) according to the original queries \( (Q) \) with high recall, (2) NER module is designed to identify candidate entities \( (E) \) of given entity type from the paragraph set \( P \), (3) Passage Ranking module is designed to return relevant passages with their ranking values which depends on the similarities between retrieved paragraphs \( (P) \) and expanded queries \( (Q') \) plus entities \( (E) \).

2.1 Paragraph Retrieval Module
This module aims to retrieve paragraph set \( (P) \) with high recall performance, thus, it would filter the irrelevant ones and reduce the burdens of the later modules. Paragraph is selected as our index unit because it contains relatively completed description on one topic. In parsing HTML document collection, we separate paragraphs according to tag <p> and record their positions in each full-text documents. And these paragraphs are pushed into index with Porter stemming applied and stop words removed. The synonyms of terms in initial queries \( (Q, 36 \text{ topics released officially}) \) are expanded to \( Q' \) based on our in-house dictionary automatically, where this expansion is only applied to specific entities (e.g., gene name \( Rb1 \)) but not to general concepts (e.g., tumor types). We collect top 4000 retrieved paragraphs from BM2500 [3] retrieval results for each topic.

2.2 NER Module
We construct an in-house biomedical dictionary from several external terminology resources including MesH (Medical Subject Headings) [5], GO (Gene Ontology) [6], SNOMED CT (Systematized Nomenclature of Medicine--Clinical Terms) [7] and UniProtKB (Universal Protein Resource Knowledgebase) [8]. Table 1 shows the resources which are applied to identify entities of given types in queries.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Entity Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeSH [5]</td>
<td>Antibody, Cell or Tissue Type, Disease, Drug, Mutation, Pathway, Signaling</td>
</tr>
<tr>
<td></td>
<td>Pathway, Tumor Type</td>
</tr>
<tr>
<td>GO Ontology [6]</td>
<td>Molecular Function</td>
</tr>
<tr>
<td>SNOMED CT [7]</td>
<td>Biological Substance, Diseases, Drug, Signs or Symptoms, Toxicities</td>
</tr>
<tr>
<td>UniProt [8]</td>
<td>Protein, Gene</td>
</tr>
</tbody>
</table>

To reduce ambiguity from different conformation of entity names, we use a series of normalize rules [4] to perform the recognition such as:

1. Prefixes and suffixes which are not critical for entity identification are removed.
2. Terms with digits or Roman/Greek numbers are transformed into a unified format.
3. Terms not in abbreviated forms are converted to lowercases.

Additionally, we use “brute force” manner to perform the recognition: forward/backword matching to get the maximum matching length.

2.3 Passage Ranking Module
This module is designed to rank candidate passages according to their relevance to the original queries. We classify terms in one query into four categories: Known Entity \( K \) (e.g., “\( Rb1 \)”), Desired Entity Type \( D \) (e.g., “tumor type”), Functional Constraints \( F \) (e.g., “associate”), and Others. Where, desired entity type is replaced by candidate entities identified in the last module. We propose the following heuristic strategies for ranking:

1. Passages with \( K \) AND \( D \) AND \( F \) occurrence should be PERFECT results;
2. Passages with \( K \) OR \( D \) AND \( F \) occurrence should be GOOD results;
3. Passages with \( K \) OR \( D \) occurrence should be NORMAL results;
4. Passages without neither \( K \) nor \( D \) occurrence should be BAD results.
We combine the above strategies with BM2500 as ranking criterion with different parameters assigned. We do the ranking process in each partition and ensure there is no exchange between different partitions, and then concatenate the four partitions together from perfect to bad and return the top 1000 results.

3. Results

Our submitted runs performed around the median compared with all the 66 official runs. To find out the bottleneck of our model, we take advantage of the official released qrels to analyze our model from different aspects: recall of paragraph retrieval module, recall of NER module and MAP of passage ranking module.

3.1 Paragraph Recall

We evaluate the recall performance of paragraph retrieval module at two different levels: document level and paragraph level. Document ID (PMID) and passage position in the official qrels are selected as golden standards to calculate Recall of retrieved documents and paragraphs as follows:

\[
\text{Doc\_Recall} = \frac{\text{Retrieved\_PMID} \cap \text{Qrels\_PMID}}{\text{Qrels\_PMID}}, \quad \text{Paragraph\_Recall} = \frac{\text{Retrieved\_Passagelength} \cap \text{Qrels\_Passagelength}}{\text{Qrels\_Passagelength}}
\]

As shown in Figure 2, the results demonstrate that paragraph retrieval model does possess promising recall (in document level 30/36 higher than 0.8 and in passage level 19/36 higher than 0.5).

3.2 Entity Recall

Answer entities in the official qrels are used as golden standard to evaluate our Recall performance of NER module:

\[
\text{NE \_Recall} = \frac{\text{Retrieved\_NE} \cap \text{Qrels\_NE}}{\text{Qrels\_NE}}
\]

As shown in Figure 3, the named entity recall performances differ from entity types, and 12/15 entity types are lower than 0.5. The major reason is that our constructed dictionary is not creditable enough to cover the entities in all the 15 types.
3.3 Passage Rank MAP
To demonstrate our passage ranking module’s capability, we use the answer entities in the official *qrels* to replace our identified candidate entities as the input of passage ranking module. Without changing any other techniques employed in our official submission, we redo the experiment and generate a run tagged as “REVISED” whose MAP performance is compared in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Document MAP</th>
<th>Aspect MAP</th>
<th>Passage MAP</th>
<th>Passage2 MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.0329</td>
<td>0.0197</td>
<td>0.0029</td>
<td>0.0008</td>
</tr>
<tr>
<td>Median</td>
<td>0.1897</td>
<td>0.1311</td>
<td>0.0565</td>
<td>0.0377</td>
</tr>
<tr>
<td>Mean</td>
<td>0.1862</td>
<td>0.1326</td>
<td>0.0560</td>
<td>0.0398</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.3286</td>
<td>0.2631</td>
<td>0.0976</td>
<td>0.1148</td>
</tr>
<tr>
<td>SUBMITTED</td>
<td>0.1603</td>
<td>0.0555</td>
<td>0.0642</td>
<td>0.0182</td>
</tr>
<tr>
<td>REVISED</td>
<td>0.2575</td>
<td>0.1543</td>
<td><strong>0.1023</strong></td>
<td>0.0444</td>
</tr>
</tbody>
</table>

Compared with our submitted runs, the revised one performs much better (Document MAP improves 60.64%, Aspect MAP improves 178.02%; Passage MAP improves 59.35% and exceeds the maximum 0.0976; Passage2 MAP improves 144%). It yields the evidences that NER module becomes the bottleneck of our multi-granularity retrieval model.

4. Conclusions
From the experiment results, we can conclude that our proposed multi-granularity retrieval model is capable to bridge the gap between biomedical concepts in queries and entities in relevant documents. Its two of three modules (paragraph retrieval and passage ranking) achieve good performance in our experiment analysis. The performance of NER module becomes the system bottleneck.

In the further work, we would focus on breaking the bottleneck caused by NER module including: (1) maintain a high quality biomedical dictionary; (2) present efficient and flexible identification methods; (3) reduce dependencies between passage ranking module with NER modules.

Acknowledgements
Thanks Chong Long and Hao Yu for their discussion during the work. We would like to thank Dr. Minlie Huang and Dr. Min Zhang for their advices.

References
7. The Systematized Nomenclature of Medicine [http://www.snomed.org/]