SNUMedinfo at CLEF QA track BioASQ 2015

Sungbin Choi

Department of Biomedical Engineering, Seoul National University, Seoul, Republic of Korea

wakeup06@empas.com

Abstract. This paper describes our participation at the BioASQ Task 3b of CLEF 2015 Question Answering track. We participated at the document retrieval subtask in Phase A and the ideal answer generation subtask in Phase B. As of previous year, in the document retrieval task, we mostly experimented with semantic concept-enriched dependence model and sequential dependence model. In the ideal answer generation task, relevant passages are selected and combined to automatically produce answer text.

Keywords: Information retrieval, Semantic concept-enriched dependence model, Sequential dependence model

1 Introduction

This paper describes the participation of the SNUMedinfo at the CLEF 2015 BioASQ task 3b. We experimented with almost similar method as of our previous participation [1].
Task 3b was about biomedical semantic question answering task. For a detailed introduction of the task, please see the overview paper of CLEF Question Answering track BioASQ 2015' [2].

2 Methods

2.1 Task 3b Phase A – Document retrieval

In Task 3b Phase A, we participated at the document retrieval subtask only. We used Indri search engine [3]. The queries are stopped at the query time using the standard 418 INQUERY stopword list, case-folded, and stemmed using Porter stemmer. We used unigram language model with Dirichlet prior smoothing [4] as our baseline retrieval method (referred as QL: query likelihood model).
We experimented with semantic concept-enriched dependence model (SCDM) [5] and sequential dependence model (SDM) [6]. For a detailed description of our retrieval method, please see our previous paper [1].

Sequential dependence model (SDM)

SDM Indri query example for the original query '*What is the inheritance pattern of Emery-Dreifuss muscular dystrophy*?' can be described as follows.

#weight (

λ_T #combine(inheritance pattern emery dreifuss muscular dystrophy)

λ_O #combine(#od1(inheritance pattern) #od1(pattern emery) #od1(emery dreifuss) #od1(dreifuss muscular) #od1(muscular dystrophy))

λ_U #combine(#uw8(inheritance pattern) #uw8(pattern emery) #uw8(emery dreifuss) #uw8(dreifuss muscular) #uw8(muscular dystrophy)))

λ_T, λ_O, λ_U are weight parameters for single terms, ordered phrases and unordered phrases, respectively.

Semantic concept-enriched dependence model (SCDM)

SCDM Indri query example can be described as follows.

- SCDM type C (single + multi-term, all-in-one)

#weight (

λ_T #combine(inheritance pattern emery dreifuss muscular dystrophy)

λ_O #combine(#od1(inheritance pattern) #od1(pattern emery) #od1(emery dreifuss) #od1(dreifuss muscular) #od1(muscular dystrophy))

λ_U #combine(#uw8(inheritance pattern) #uw8(pattern emery) #uw8(emery dreifuss) #uw8(dreifuss muscular) #uw8(muscular dystrophy))

λ_{O_SC} #combine(#od1(inheritance pattern) #od1(emery dreifuss muscular dystrophy))

λ_{U_SC} #combine(#uw8(inheritance pattern) #uw16(emery dreifuss muscular dystrophy)))

λ_T, λ_O, λ_U, λ_{O_SC}, λ_{U_SC} are weight parameters for single terms, ordered phrases and unordered phrases of sequential query term pairs, ordered phrases and unordered phrases of semantic concepts, respectively.

- SCDM type D (single+multi-term, pairwise)

#weight (

λ_T #combine(inheritance pattern emery dreifuss muscular dystrophy)

λ_O #combine(#od1(inheritance pattern) #od1(pattern emery) #od1(emery dreifuss) #od1(dreifuss muscular) #od1(muscular dystrophy))

λ_U #combine(#uw8(inheritance pattern) #uw8(pattern emery) #uw8(emery dreifuss) #uw8(dreifuss muscular) #uw8(muscular dystrophy))

λ_{O_SC} #combine(#od1(inheritance pattern) #od1(emery dreifuss) #od1(dreifuss muscular) #od1(muscular dystrophy))

λ_{U_SC} #combine(#uw8(inheritance pattern) #uw8(emery dreifuss) #uw8(dreifuss muscular) #uw8(muscular dystrophy)))

We experimented with following parameter settings.

SNUMedinfo1: SCDM Type C (mu=500, λ_T=0.85, λ_O=0.00, λ_U=0.00, λ_{O_SC}=0.10, λ_{U_SC}=0.05)
SNUMedinfo2: SCDM Type C (mu=500, λ_T=0.70, λ_O=0.00, λ_U=0.00, λ_{O_SC}=0.20, λ_{U_SC}=0.10)
SNUMedinfo3: SCDM Type C (mu=500, λ_T=0.70, λ_O=0.10, λ_U=0.05, λ_{O_SC}=0.10, λ_{U_SC}=0.05)
SNUMedinfo4: SCDM Type D (mu=500, λ_T=0.85, λ_O=0.00, λ_U=0.00, λ_{O_SC}=0.10, λ_{U_SC}=0.05)
SNUMedinfo5: SDM (mu=500, λ_T=0.85, λ_O=0.00, λ_U=0.00, λ_{O_SC}=0.10, λ_{U_SC}=0.05)

2.2 Task 3b Phase B – Ideal answer generation

In Task 3b Phase B, we participated only at the ideal answer generation subtask. We reformulated this task as, among relevant lists of passages given[1], selecting most appropriate ones. We experimented with following heuristic method to select m passages and combine them to form the ideal answer.

Identifying keyword terms and rank passages based on the number of unique keywords it contain

Firstly, candidate passages are ranked based on number of keywords. Parameter *minDF* represents minimum proportion of passages that keyword term should occur. If there are 20 relevant passages given, and *minDF* is set to 0.5, then any terms occurring ≥ 10 passages are considered as keywords. With identified keywords list, we rank passages based on the number of unique keywords each passage contains.

Then, passages from top ranked ones are included for answer generation. Parameter *minUnseen* represents minimum proportion of new tokens that does not exist in the previously selected passages. We check proportion of tokens in the passage that does not occur in the previously selected passages, and if it is ≥ *minUnseen* threshold, second-ranked passage is selected. If proportions of newly found tokens are below *minUnseen* threshold, that passage is abandoned, and we check next rank passage. This process is repeated until *m* passage is selected. We intend to enhance comprehensiveness of answer text by increasing the diversity of tokens.

[1] We used gold relevant text snippets provided by the BioASQ.

In this method, our intention was enhancing comprehensiveness of answer text by increasing the diversity of tokens.

3 Results & Discussion

At the moment of writing this paper, the final evaluation results are not available yet. So we report tentative evaluation results currently available for us.

3.1 Task 3b Phase A – Document retrieval

There were five distinct batches within this task.

Table 1. Tentative evaluation results of submitted runs (Evaluation metric: MAP)

	SNU Medinfo1	**SNU Medinfo2**	**SNU Medinfo3**	**SNU Medinfo4**	**SNU Medinfo5**
Batch1	0.1733	0.1731	0.1695	0.1724	0.1569
Batch2	0.2250	0.2229	0.2205	0.2245	0.2111
Batch3	0.2022	0.2015	0.2089	0.1973	-
Batch4	0.1647	0.1625	0.1728	0.1650	0.1653
Batch5	0.1772	0.1794	0.1772	0.1765	0.1890

Generally, SDM and SCDM showed better performance compared to the baseline QL method. But compared to the previous year, limit of returned document per query is decreased from 100 to 10. We presume that the evaluation scores become more volatile because of that.

3.2 Task 3b Phase B – Ideal answer generation

We submitted five runs trying different parameter values, but according to the automatic evaluation score (Rouge-2 and Rouge-SU4) evaluation, performance change seems not very meaningful.

Table 2. Tentative evaluation results of submitted runs (Evaluation metric: ROUGE-SU4)

	SNU Medinfo1	**SNU Medinfo2**	**SNU Medinfo3**	**SNU Medinfo4**	**SNU Medinfo5**
Batch1	0.3069	0.3071	0.3034	0.2703	0.2784
Batch2	0.3597	0.3710	0.3742	0.3268	0.3461
Batch3	0.3950	0.3941	0.3906	0.3690	0.3754
Batch4	0.3684	0.3906	0.3644	0.3439	0.3556
Batch5	0.3532	0.3665	0.3484	0.3202	0.3282

4 References

1. Choi, S. and J. Choi, Classification and retrieval of biomedical literatures: Snumedinfo at clef qa track bioasq 2014. Proceedings of Question Answering Lab at CLEF, 2014.
2. Cappellato, L., Ferro, N., Jones, G., and San Juan, E., CLEF 2015 Labs and Workshops. 2015, CEUR Workshop Proceedings (CEUR-WS.org).
3. Strohman, T., et al. Indri: A language model-based search engine for complex queries. in Proceedings of the International Conference on Intelligent Analysis. 2005. McLean, VA.
4. Zhai, C. and J. Lafferty, A study of smoothing methods for language models applied to Ad Hoc information retrieval, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001, ACM: New Orleans, Louisiana, USA. p. 334-342.
5. Choi, S., et al., Semantic concept-enriched dependence model for medical information retrieval. Journal of Biomedical Informatics. **47**: p. 18-27.
6. Metzler, D. and W.B. Croft, A Markov random field model for term dependencies, in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005, ACM: Salvador, Brazil. p. 472-479.