The Fudan participation in the 2015 BioASQ Challenge: Large-scale Biomedical Semantic Indexing and Question Answering

Shengwen Peng[1,2], Ronghui You[1,2], Zhikai Xie[1,3], Beichen Wang[1,2], Yanchun Zhang[1,3,4], and Shanfeng Zhu[1,2] ⋆

[1] School of Computer Science, Fudan University, Shanghai 200433, P. R. China,
[2] Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, P. R. China
[3] Shanghai Key Lab of Data Science, Fudan University, Shanghai 200433, P. R. China
[4] Centre for Applied Informatics, College of Engineering and Science, Victoria University, Australia
{14210240017,11300720164,13210240118,12210240068,yanchunzhang,zhusf}@fudan.edu.cn

Abstract. This article describes the participation of Fudan team in the 2015 BioASQ challenge. The challenge consists of two tasks, large-scale biomedical semantic indexing (task 3a) and biomedical question answering (task 3b). In task 3a, our method, MeSHLabeler, achieved the first place in all 15 weeks of three batches. Based on 3215 annotated citations (June 6, 2015) out of all 4435 citations in the official test set (batch 1, week 2), our submission best achieved 0.6194 in flat Micro F-measure. This is 0.0576(10.25%) higher than 0.5618, obtained by the official NLM solution Medical Text Indexer (MTI). Task 3b includes two phases. Given the questions raised by a team of biomedical experts from around Europe, the main task of phase A is to find relevant documents, snippets, concepts and RDF triples, while the main task of phase B is to provide exact and ideal answers. In the phase A of task 3b, our submission, fdu, achieved the first place in both document and snippet retrieval in batch 5 (June 6, 2015).

Keywords: MeSH Indexing; Learning to Rank; Multi-Label Classification; Biomedical Question Answering; Information Retrieval; Information Extraction.

1 Introduction

BioASQ 2015 is the third year of BioASQ challenge, which is an established international competition for large-scale biomedical semantic indexing and question answering since 2013[1]. It consists of two tasks, A) automatic indexing new MEDLINE citations using current Medical Subject Headings (MeSH), and B)

⋆ Corresponding author

answering biomedical questions raised by biomedical expert from around Europe. Task 3a includes three rounds, with each taking five weeks. In each week, the organizers provide thousands of new MEDLINE citations to the competition participants, who are required to submit MeSH annotations in 21 hours. Our system, MeSHLabeler, achieved the first place in all 15 weeks of three rounds. Task 3b includes 5 rounds. In each round, around 100 biomedical questions are provided to the competition participants. There are two phases in each round. In phase A, the competition participants are required to submit relevant documents, snippets, concepts and RDF triples in 24 hours. The organizers then release gold (correct) relevant articles and snippets. In phase B, the competition participants are required to submit exact and ideal answers of these questions. In the phase A of task 3b, our best submission, fdu, achieved the first place in both document and snippet retrieval in round 5, with MAP of 0.2035 and 0.1226, respectively.

2 Task3a: Large-scale MeSH Indexing

2.1 Problem

MeSH [2] is the NLM controlled vocabulary thesaurus used for indexing almost all of the citations in MEDLINE[5]. It is widely used for facilitating biomedical information retrieval and knowledge discovery[3, 4, 5, 6]. MeSH is organized according to the hierarchical structure, and it is slightly updated every year. In 2015, there are more than 27000 MeSH headings. With the dramatic growth of biomedical documents, the number of citations in MEDLINE has reached to 23 million [6]. To reduce the time and financial cost, NLM has developed a software package, MTI (Medical Text Indexer), for assisting MeSH annotation[7, 8, 9]. Automatic MeSH indexing is a very challenging problem. The difficulty of this problem comes from the following three factors: (1) the large number of distinct MeSH and their uneven distribution in the MEDLINE; (2) the large variation in the number of MeSH of each citation; and (3) insufficient information, such as full text.

2.2 MeSHLabeler

Each MeSH can be viewed as a label. The MeSH indexing problem is actually a large-scale multi-label classification problem. We have developed a novel algorithm, MeSHLabeler, for solving this problem, which has also achieved the first place in the round 3 of BioASQ 2014 Task2a [10, 11]. The basic idea of MeSHLabeler is to integrate multiple evidence by learning to rank to achieve accurate MeSH annotation. It consists of two components, MeSHRanker and MeSHNumber. Given a test citation x, MeSHRanker returns an ordered list L of candidate MeSH headings, and MeSHNumber predicts the actual number of

[5] http://www.ncbi.nlm.nih.gov/mesh

[6] http://www.nlm.nih.gov/pubs/factsheets/medline.html

MeSH headings annotated, K. Then top K MeSH headings of L is returned as the predicted MeSH annotation for x.

Multiple evidence has been integrated into MeSHLabeler by learning to rank. These evidence can be mainly divided into five different types, local evidence, global evidence, pattern matching, MTI and MeSH dependency. Local evidence considers only a small number of citations that are most similar to the test citation. Global evidence refers to the MeSH classifiers trained from the whole MEDLINE collection. We train a distinct classifier by logistic regression for each MeSH heading. A novel score normalization method is developed to make the prediction scores of different classifiers comparable. Pattern matching tries to scan the title and abstract of test citation to see if it includes any MeSH headings and its synonyms. MTI is a mixture of local evidence, pattern matching and indexing rules. Incorporating MTI into MeSHLabeler can take advantage of domain knowledge in MeSH indexing. In addition, taking MeSH dependency into consideration is a distinct feature of MeSHLabeler, which has not been explored in previous studies. Please refer to [11] for the detailed description of MeSHLabeler.

2.3 Data Processing and Implementation

We downloaded the entire database of MEDLINE in Feb 2015, including 23,343,329 citations. After removing the citations without abstract, title or MeSH annotations, there are 13,156,128 citations. We extracted the latest 20,000 citations as the test set and validation set of logistic regression classifiers. In addition, L-TR(Learning to Rank) dataset were extracted from the latest annotated citations during the competition. The system was mainly written in C++. Referenced external libraries includes LibLinear [7] for Logistic Regression [12], RankLib [8] for LambdaMART, JsonCpp [9] for the input/output of json files and OpenMP [10] for parallel processing. Our server has 4 * Intel XEON E5-4650 2.7GHzs CPU, and 128G memory. After extracting features, both training LTR model and making prediction are very quick. It takes 5 days to train 27,000 MeSH classifiers by Logistic Regression, but less than 2 hours to annotate 10,000 citations.

2.4 Performance

The main evaluation metrics for BioASQ are label-based micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F)[19]. During the whole competition, MeSHLabeler kept the first place on both metrics. As shown in Table 1, we compare the performance of MeSHLabeler with two baselines of BioASQ 2015, NCBI system(MeSH Now) and the MTI system, in batch 3 and week 5 of batch 2. In the week 5 of batch 2, MTI is not incorporated into the

[7] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[8] http://sourceforge.net/p/lemur/wiki/RankLib/

[9] http://jsoncpp.sourceforge.net/

[10] http://openmp.org/wp/

MeSHLabeler. MeSHLabeler achieved an MiF of 0.6247, which is 0.0426(7.3%) higher than 0.5821, obtained by MTIDEF. On the other hand, by incorporating MTI into MeSHLabeler in batch 3, the performance of MeSHLabeler is significantly improved, which is on average 8.9% higher than MTI in terms of MiF. From this we can clearly see the advantage of integrating diverse evidence, such as MTI, in MeSHLabeler for the accurate MeSH indexing. As shown in Table 2, we further compare the performance among MeSHLabeler, AUTH and MeSH-UK systems, which are the top 3 systems in the last batch.

	MiF				LCA-F			
Round	MeSHLabeler	MeSH Now	MTIDEF	MTIFL	MeSHLabeler	MeSH Now	MTIDEF	MTIFL
Batch2 Week5	**0.6247**	0.5984	0.5821	0.5807	**0.5092**	0.4958	0.4850	0.4797
Batch3 Week1	**0.6339**	0.6036	0.5836	0.5782	**0.5198**	0.5070	0.4945	0.4833
Batch3 Week2	**0.6399**	0.6022	0.5841	0.5690	**0.5244**	0.5042	0.4923	0.4721
Batch3 Week3	**0.6445**	0.6004	0.5920	0.5737	**0.5316**	0.4995	0.5026	0.4811
Batch3 Week4	**0.6267**	0.5929	0.5690	0.5563	**0.5157**	0.4989	0.4829	0.4672
Batch3 Week5	**0.6358**	0.5971	0.5832	0.5725	**0.5198**	0.4971	0.4905	0.4791

Table 1. The performance comparison of MeSHLabeler, NCBI(MeSH Now) and MTI in batch 3 and week5 of batch 2 (Updated on July 6, 2015)

	MiF			LCA-F		
Round	MeSHLabeler	AUTH	MeSH-UK	MeSHLabeler	AUTH	MeSH-UK
Batch2 Week5	**0.6247**	0.6012	0.6140	**0.5092**	0.4958	0.4992
Batch3 Week1	**0.6339**	-	0.6205	**0.5198**	-	0.5080
Batch3 Week2	**0.6399**	0.6235	0.6196	**0.5244**	0.5101	0.5081
Batch3 Week3	**0.6445**	0.6239	0.6105	**0.5316**	0.5144	0.5019
Batch3 Week4	**0.6267**	0.6104	0.6059	**0.5157**	0.5053	0.4919
Batch3 Week5	**0.6358**	0.6148	0.6103	**0.5198**	0.5083	0.4933

Table 2. The performance comparison of MeSHLabeler, AUTH and MeSH-UK in batch 3 and week5 of batch 2 (Updated on July 6, 2015)

3 Task 3b: Biomedical Question Answering

There are 4 types of questions in task 3b: yes/no, factoid, list and summary. In the phase A of task 3b, the participants are required to submit the lists of relevant documents, concepts, snippets and RDF triples. In each list, at most 10 relevant items can be returned. In the phase B of task 3b, the participants are required to return the exact and ideal answers.

3.1 Task 3b Phase A: Find relevant documents and snippets

Here we briefly describe several important factors considered in our retrieval system.

Retrieval Model We chose a statistical language model [13], query likelihood model, as the underlying model to retrieve relevant documents. The open source software package Indri [11] was used to build our system, which achieves good performance in many applications.

Term Weight Optimization As most of nouns in the query are concepts and phrases, we think these keywords are more important than other keywords. To emphasize these keywords, we put higher weights on these keywords in the model.

Occurrence of Query Keywords We check the occurrence of query keywords in the retrieved documents. A document that includes all keywords in the query should be scored higher than other documents with very few query keywords.

Stemming We find that stemming greatly affects the performance of the retrieval system. In most cases, no stemming is a better choice. However, for some questions, stemming could improve the retrieval performance.

Phrase If two terms are adjacent in the query, they may be a part of phrase. In this case, the documents that includes the phrase should be emphasized.

Pseudo Relevance Feedback In our information retrieval, we use the pseudo relevance feedback to select the top k documents. Some keywords in these documents are used for query expansion.

3.2 Task 3b Phase B: Provide exact and ideal answer

Since the golden relevant documents, concepts, snippets and RDF triples are available for the participants in phase B, we use these to extract exact answers and ideal answers. After checking previous work of other teams [14, 15, 16], we design our question answering system. As shown in Fig. 1, our system architecture of question answering consists of three main components, question analysis, candidates generating, and candidates ranking.

[11] http://www.lemurproject.org/indri.php

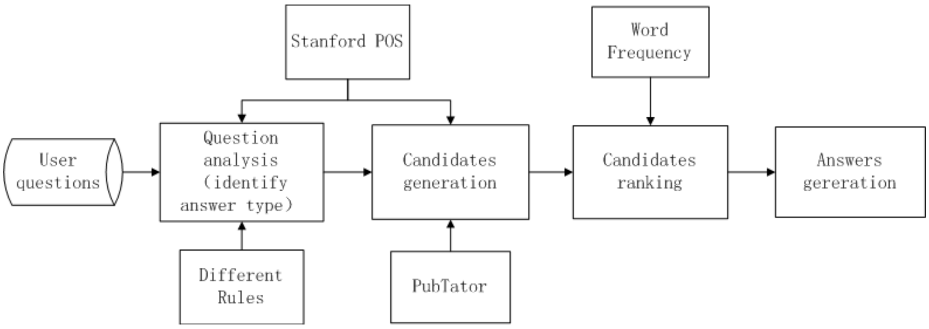

Fig. 1. The system architecture of QA

Question analysis Question analysis is mainly responsible for extracting answer types of questions. It is important to understand what the question is actually asking about. For Factoid and List-type questions, they usually carry key information of answer type. We classify questions into several types of desired answers: 1) disease; 2) drug; 3) gene/protein; 4) mutation; 5) number; 6) choice. Based on the above strategies and historical data of BioASQ 2013 and BioASQ 2014 [17], we develop a set of rules to recognize answer types of given questions. For example, A Factoid question is as follows: Which gene is associated with Muenke syndrome? The corresponding rule is "Which gene (.)*".

Candidates generating We could use different methods to generate candidates of different questions. For diseases, drugs, gene, protein, mutation, and other biomedical questions, we can employ PubTator [18] tool to identify and extract biomedical concepts of corresponding answer type as candidates. For number and other questions, we use Stanford POS tool [12] to tag relevant snippets, noun, noun phrases and numbers as candidates.

Candidates ranking For each candidate, we count its word frequency in relevant documents and snippets, which is the basis of ranking. We then return the maximum number of allowed answers (e.g., no more than 100 answers for List-type question).

3.3 Performance

In Phase A of task 3b, our best submission achieved the first place in finding relevant documents and snippets in batch 5, with a MAP score of 0.2035 and 0.1226, respectively. The performance of fdu and the best MAP of SNUMedinfo in document level is shown in Table 3. Moreover, Table 4 illustrates the best performance of top 3 teams in terms of exact answer on task 3b Phase B (July

[12] http://nlp.stanford.edu/software/tagger.shtml

6, 2015). According to official measures of different types of questions, we obtain the best performance for the Yes/No-type questions in the Batch 5, and for the List-type questions in Batch 2. Overall our system achieved good performance in all three types of questions in all five batches.

Batch	Mean precision	Recall	F-Measure	MAP	GMAP	SNUMedinfo(MAP)
1	0.2320	0.3275	0.2232	0.1719	0.0071	**0.1733**
2	0.2990	0.3711	0.2730	**0.2264**	0.0217	0.2250
3	0.2530	0.3378	0.2515	0.1762	0.0154	**0.2089**
4	0.2196	0.3498	0.2353	0.1597	0.0080	**0.1728**
5	0.2640	0.5270	0.3194	**0.2053**	0.0290	0.1890

Table 3. The document level performance of fdu and the best MAP of SNUMedinfo in all 5 batches of Task 3b Phase A

4 Discussion and Conclusion

Although MeSHLabeler achieved great performance in automatic MeSH indexing, there are several issues to be explored in the future. Firstly, we found some citations lack of similar citations from Entrez elink. The data inconsistency can lead to bad prediction accuracy for these citations. Considering the importance of KNN score, we may find similar citations in other ways for these citations. Secondly, we did not take advantage of the MeSH hierarchical structure, which could be used to optimize the LCA-F score. Finally, although we attempted to integrate other information into MeSHLabeler, the performance varies only slightly. We would like to know the upper limit of MeSHLabeler and if we can find an effective method to integrate other types of information.

For the Phase A of task 3a, we find that stemming sometimes improves the performance. The interesting future work is to automatically judge whether stemming is a good choice for a specific question. For the Phase B of task 3b, we make use of PubTator to identify biomedical concepts. However, some biomedical concepts cannot be recognized. An accurate biomedical concept identification tool with high coverage would be important for the success of biomedical question answer system.

5 Acknowledgement

This work has been partially supported by National Natural Science Foundation of China (61332013, 61272480 and 61170097), and Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China. We would like to thank Ke Liu, Junning Gao and Renyu Liu for their helpful suggestions and insightful discussion.

Batch	Yes/No(Accuracy)	Factoid(MRR)	List(F-Measure)
1	**fa1(0.8458)**	**main system(0.1938)**	**main system(0.1362)**
	fdu(0.8458)	fdu(0.1423)	fdu(0.0756)
	main system(0.8458)	fa1(0.0769)	HPI-S2(0.0650)
2	**main system(0.8125)**	**main system(0.1198)**	**fdu(0.1160)**
	fa1(0.8125)	fdu (0.0859)	main system(0.1081)
	fdu(0.8125)	fa1(0.0313)	HPI-S2(0.0262)
3	**main system(0.9655)**	**oaqa-3b-3(0.1615)**	**main system(0.1587)**
	fa1(0.9655)	main system(0.1346)	fdu(0.1319)
	fdu(0.9655)	fdu(0.0846)	oaqa-3b-3-e(0.0969)
4	**oaqa-3b-4(0.9600)**	**oaqa-3b-4(0.5155)**	**oaqa-3b-4(0.3168)**
	main system(0.9600)	main system(0.3201)	fdu(0.2192)
	fdu(0.9600)	fdu(0.2299)	main system(0.1349)
5	**fdu(0.7143)**	**oaqa-3b-4(0.2727)**	**oaqa-3b-5(0.1875)**
	fa1(0.6786)	fdu(0.2500)	YodaQA_base(0.1631)
	main system(0.6786)	YodaQA_base(0.2045)	fdu(0.1340)

Table 4. The performance of top 3 systems over exact answer in all 5 batches of Task 3b Phase B

References

[1] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., et al. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics, 16(1), 138.*

[2] Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J. L., & Arluk, N.: The MeSH translation maintenance system: structure, interface design, and implementation. Medinfo, 11(Pt 1), 67–69. (2004)

[3] Gu, J., Feng, W., Zeng, J., Mamitsuka, H., Zhu, S.: Efficient Semi-supervised MEDLINE document clustering with MeSH semantic and global content constraints. IEEE Transactions on Cybernetics, 43 (4), 1265–1276, (2013)

[4] Zhu, S., Zeng, J., Mamitsuka, H.: Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics 25(15): 1944–1951 (2009)

[5] Zhu, S., Takigawa, I., Zeng, J., Mamitsuka, H.: Field independent probabilistic model for clustering multi-field documents. Information Processing & Management. 45(5): 555–570 (2009)

[6] Huang, X., Zheng, X., Yuan, W., Wang, F., Zhu, S.: Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. Information Science 181(11): 2293–2302 (2011)

[7] Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R.: The NLM Medical Text Indexer System for Indexing Biomedical Literature. In BioASQ@ CLEF. (2013)

[8] Aronson, A., Mork, J., Gay, C., Humphrey, S., and Rogers, W. Then NLM indexing initiative's medical text indexer. Stud Health Technol Inform. 107(Pt 1). 268-272

[9] Mork, J., Demner-Fushman, D., Schmidt, S., and Aronson, A. Recent Enhancements to the NLM Medical Text Indexer. CLEF (Working Notes) 2014: 1328-1336

[10] Liu, K., Wu, J., Peng, S., Zhai, C., Zhu, S.: The Fudan-UIUC Participation in the BioASQ Challenge Task 2a: The Antinomyra system. CLEF (Working Notes) 2014: 1311-1318

[11] Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka H., Zhu, S. (2015).: MeSHLabeler: Improving the Accuracy of Large-scale MeSH indexing by Integrating Diverse Evidence. Bioinformatics 31(12): 339-347 (2015)

[12] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., Lin, C. J.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research, 9: 1871–1874. (2008)

[13] Zhai, C. Statistical Language Models for Information Retrieval: A Critical Review at Foundations and Trends in Information Retrieval 2(3): 137-213 (2008)

[14] Weissenborn, D., Tsatsaronis, G., Schroeder, M. (2013). Answering factoid questions in the biomedical domain. In 1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering.

[15] Mao, Y., Wei, C.H., Lu, Z. (2014). NCBI at the 2014 BioASQ challenge task: large-scale biomedical semantic indexing and question answering.

[16] Choi, S., Choi, J. (2014). Classification and Retrieval of Biomedical Literatures: SNUMedinfo at CLEF QA track BioASQ2014.

[17] Balikas, G., Partalas, I., Ngomo, A. C. N., Krithara, A., Gaussier, E., & Paliouras, G. (2014). Results of the BioASQ Track of the Question Answering Lab at *CLEF 2014. CLEF.*

[18] Wei, C.-H., Kao, H.-Y., Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research 41, W518-W522.

[19] Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., & Androutsopoulos, I. (2013). Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining and Knowledge Discovery, 29(3), 820-865.