NCBI at the 2014 BioASQ challenge task: large-scale biomedical semantic indexing and question answering

Yuqing Mao[1], Chih-Hsuan Wei[1], Zhiyong Lu[1,*]

[1] National Center for Biotechnology Information (NCBI),
8600 Rockville Pike, Bethesda, MD 20894, USA

{yuqing.mao,chih-hsuan.wei,zhiyong.lu}@nih.gov

Abstract. In this paper we report our participation in the 2014 BioASQ challenge tasks on biomedical semantic indexing and question answering. For the biomedical semantic indexing task (Task 2a) where participating teams are provided with PubMed articles and asked to return relevant MeSH terms, we built on our previous learning-to-rank framework with a special focus on systematically incorporating results of complementary methods for improved performance. For the question answering task (Task 2b) where teams are provided with natural language questions and asked to return responses in the format of documents, snippets, concepts and RDF triplets (Phase A) and direct answers (Phase B), we relied on PubMed search engines and our state-of-the-art named entity recognition tools such as DNorm and tmVar in Phases A and B, respectively. The official challenge results demonstrate that we consistently performed better than the baseline approaches for Task 2a and Task 2b (Phase B), and ranked among the top tier systems in the 2014 challenge.

Keywords: MeSH Indexing; Biomedical Semantic Indexing; Hierarchical Text Classification; Learning to Rank; Biomedical Question Answering.

1 Introduction

Over the past decade, a number of community-wide challenge evaluations have been held for various research topics in the biomedical natural language processing (BioNLP) field, such as document retrieval [1, 2], named entity recognition [3-5], information extraction [6, 7], etc. Different from other challenges such as BioCreative [8, 9], the BioASQ Challenge (http://www.bioasq.org/) is a newly organized shared task and has a unique focus on biomedical semantic indexing and question answering.

Similar to the previous year [10], the BioASQ 2014 challenge consists of two tasks: automated semantic (MeSH) indexing (Task 2a) and question answering (Task 2b). More specifically, for Task 2a, participating teams are provided with a set of newly published articles in PubMed, and are asked to automatically predict the most relevant MeSH terms for each article in the given set. For evaluation, team prediction results will be compared to those gold standard curated by human indexers. MeSH indexing

is an important task for the US National Library of Medicine (NLM) because indexed MeSH terms can then be used implicitly or explicitly for searching the biomedical literature in PubMed [11]. Indexed MeSH terms can also play a role in many other scientific investigations [12-14] in the biomedical informatics research.

However, like many other curation tasks, manual MeSH indexing is labor-intensive and time-consuming. As shown in [15, 16], it can take weeks or even months for an article to be manually indexed with relevant MeSH terms after it first enters PubMed. In response, many automated systems for assisting MeSH indexing have been proposed in the past [15-17]. Some automated systems such as the NLM Medical Text Indexer (MTI) and its newer version, Medical Text Indexer First Line (MTIFL) [18], are already being used in the NLM production pipelines to assist human annotators with indexing MeSH main headings, and main heading/subheading pairs [19].

Task2b is a biomedical question-answering task. For this task, teams are provided with 100 natural language questions in each batch (5 test batches in total) and asked to return answers in two phases. In Phase A, the participating teams should return relevant documents, concepts, RDF triples and snippets for each question. In Phase B, the teams should return "exact" and "ideal" answers. Exact answers depend on the question type, which can be categorized as below:

- Yes/no type questions: answer either yes or no
- Factoid type questions: answer named entities
- List type questions: answer list of named entities
- Summary type questions: no exact answer is needed

Ideal answers are paragraph-sized summaries that are required for all four types of questions. For both phases of Task 2, the question type is known to the participants.

2 Methods

2.1 Task 2a

For Task 2a, our overall approach builds on our previous research where we first proposed to reformulate the MeSH prediction task as a ranking problem in 2010 [16]: our approach first retrieves an initial list of MeSH terms as candidates for each target article. Next, we apply a learning-to-rank algorithm to re-rank the candidate MeSH terms based on the learned associations between the document text and each candidate MeSH term. More specifically, each main heading (MH) candidate can be represented as a feature vector as $x_i = (x_1^i, x_2^i, \ldots, x_m^i)$, where m is the number of features (e.g. neighborhood features, unigram/bigram features, etc). The learning objective is to find a ranking function $f(x)$ which can assign a score to each main heading based on the feature vector and subsequently use the scores to rank relevant main headings of

the target document ahead of those irrelevant ones. Finally, we prune the ranked list and return a number of top candidates as the final system output.

Through participation in the indexing task in BioASQ 2013 [20], we have shown several useful extensions such as using a different learning-to-rank algorithm with an enriched set of learning features, as well as using different methods for list-pruning and selecting top candidates from the ranked list.

In BioASQ 2014, we further expanded our approach in the following aspects: First, we built binary SVM classifiers using bag-of-word features, one for each MeSH term, as suggested by [21]. Second, predicted MeSH terms from the aforementioned binary classifiers and NLM's MTI system were added to our list of candidate MeSH terms, in addition to those already collected from the neighbor documents. Third, we limited the neighbor documents to newly indexed articles (last five years) and used a more recent and larger training set, along with a new list-pruning method for selecting final terms from our ranked list. Lastly, we used some post-processing techniques, such as using string matching to identify "Age Check Tags" in the abstract, to enhance the final system output. Table 1 shows a detailed list of key differences between our current system and our 2013 system. In addition to the abovementioned differences, the table also includes a few other notable modifications such as upgrading our lexicon to MeSH 2014 version.

Table 1. Major Differences between our current work and our previous approach in BioASQ 2013 (In both cases, the general learning-to-rank framework [16] was used)

Notable Differences	BioASQ 2013	BioASQ 2014
Learning-to-rank algorithm	MART	LAMBDA-MART
Neighbor documents	Retrieved from all MEDLINE database	Retrieved from documents indexed after 2009
The list of candidate MeSH terms	All MeSH terms in neighbor documents	All MeSH terms in neighbor documents plus MTI and binary classifier results
Features used in the learning-to-rank algorithm	All features in [16] plus a new feature representing the MTI results.	All features in [16] plus two new features representing the MTI and binary classifier results
MeSH version	MeSH 2013	MeSH 2014
Training data for the learning-to-rank algorithm	1000 documents from select BioASQ Journal List	5000 documents from select BioASQ Journal List
Method for selecting the number of predicted MeSH terms from the ranked list	$S_{i+1} / S_i < i/(i+1+\alpha)$ S_i is the score of predicted MeSH Term at position *i*	$S_{i+1} < S_i \cdot \log(i) \cdot \lambda$ S_i is the score of predicted MeSH Term at position *i.*
Post-process techniques	NONE	Refine Age Check Tags Add tags like "Europe" to European foreign journals

2.2 Task 2b – Phase A

For returning relevant documents, we used PubMed search functions. Given a search query, PubMed provides users with two results-ranking options: by date or by relevance. Furthermore, we computed cosine similarity (Eq.1) scores between the question (q) and each sentence (s) in a retrieved article. The sentence in the abstract with the highest score would be returned as a snippet. We did not use full text in this work.

$$\cos(q, s) = \frac{q \cdot s}{\|q\| \|s\|} = \frac{\sum_{t \in q \cap s} q_t \cdot s_t}{\sqrt{\sum {q_t}^2} \sqrt{\sum {s_t}^2}} \qquad (1)$$

For concept recognition, we used a dictionary-look up method to mine disease, chemical and GO terms and used our previous developed gene normalization tool, GenNorm [22], to identify gene/protein mentions. In addition, we used MetaMap [23] to extract MeSH concepts from the questions. For snippets, we only return results when the relevant concepts are gene/proteins.

2.3 Task 2b – Phase B

In phase B, the gold-standard relevant documents, concepts, snippets and RDF triples in Phase A become available to the participants. In particular, we used the relevant documents and snippets for returning "exact" answers in Phase B.

"Exact" answers: For "yes/no" type questions, we simply returned "yes" as "exact" answers because of its strong performance on the previous training data. No "exact" answers were needed for summary-type questions.

For Factoid and List-type questions, we developed a three-step approach for returning "exact" answers. The first step was to automatically determine the type of desired answers: 1) numbers; 2) multiple choices; and 3) bio-concepts. (See examples in Table 2). If bio-concepts are desired, we further classified them into sub-types: 3a) Gene/proteins; 3b) Chemical/drugs; 3c) Disorder/syndromes; 3d) Mutation/variations and 3e) Species/viruses. Based on such a strategy and previous year's data, we developed a set of regular expression patterns to identify different answer types and subtypes for a given question. When no match is found, the question will be discarded from further processing (i.e. no "exact" answers will be returned). Otherwise, it will be passed to the next step.

The next step was to generate candidate answers for different answer types in Factoid and List-type questions. For 1), we identified all numbers in relevant snippets to be candidate answers. For 2), the candidates were mined from the questions. For 3) we applied our PubTator [24-26] tool to the relevant documents for obtaining entity recognition results when generating the candidate answers. PubTator is equipped with several competition-wining text-mining algorithms for automatically extracting bio-concepts (GenNorm [22] for genes, tmChem [4] for chemicals, DNorm for [27] diseases, SR4GN [28] for species, and tmVar [29] for mutations) from free text.

Table 2. Three answer types for Factoid and List-type questions.

Answer Type	Example questions
1) Numbers	How **many** genes does the human hoxD cluster contain? What is the **incidence** of Edwards's syndrome in the European population?
2) Multi-Choices	Is the transcriptional regulator BACH1 an **activator** or a **repressor**?
3) Bio-concepts	Which **gene** is involved in CADASIL? Which **drugs** affect insulin resistance in obesity? Which **disease** is caused by mutations in Calsequestrin 2 (CASQ2) gene? Which gene **mutations** are responsible for isolated Non-compaction cardiomyopathy? Which **virus** is Cidofovir (Vistide) indicated for?

The last step was to rank the candidate answers. For each candidate, we calculated its cosine similarity score against the relevant snippets, and ranked the candidates by the similarity scores. We then returned the maximum number of allowed answers (e.g., no more than 5 answers for Factoid-type questions).

"Ideal" answers: For returning "ideal" answers, we used the same method as retrieving relevant snippets in Phase A. That is, we scored each gold-standard snippet against the question using cosine similarity and returned the one with the highest score. This method is applied to all questions regardless of their types.

3 Results

3.1 Task 2a

Task 2a was organized for three consecutive periods (batches) of five weeks each. Each week, participants have a limited response time (less than one day) to return their predicted MeSH terms for a set of newly indexed articles in PubMed.

For Task 2a, team results were evaluated based on multiple measures. Two main measures are: the flat measure "label-based micro F-measure (MiF)" and the hierarchical measure "Lowest Common Ancestor F-measure (LCA-F)" [30].

Table **3** shows our best results on the Task 2a Batch 3 Week 2 test set, which contains the largest number of test articles (5,717) with known answers among all 15 test sets of Task 2a, as of June 30, 2014. In this run, we incorporated both MTI and binary classifier results, and applied all three post-processing methods in Table 1.

According to the official website, we rank the first in both the flat (MiF) and hierarchical F-measure (LCA-F) on this test set. We also obtained the highest recall scores in both flat and hierarchical measures (MiR and LCA-R).

Table 3. Official results (as of June 30, 2014) for our best run (L2R-n2) on Batch 3 Week 2 test set plus the results of several baseline methods. Our best results among all team submissions are highlighed in bold.

Systems	MiF	MiP	MiR	LCA-F	LCA-P	LCA-R
NCBI (L2R-n2)	**0.6052**	0.6191	**0.5919**	**0.5105**	0.5324	**0.5208**
Default MTI	0.5640	0.5921	0.5385	0.4834	0.5245	0.4770
MTI First Line	0.5520	0.6257	0.4939	0.4686	0.5434	0.4382
BIoASQ Baseline	0.2666	0.2413	0.2978	0.3120	0.3224	0.3299

3.2 Task 2b – Phase A

The test dataset of Task 2b was released in five batches[1] over a period of three months, each containing 100 questions. Several measures were used to evaluate team submissions. Table 4 shows our submission for the final batch (fifth batch) according to the official results released on June 30, 2014, where we obtained the best F-measure and mean precision for returning relevant concepts (highlighted in bold).

Table 4. Official results for our best submssion on Batch 5 Phase A test set. Our best results among all submissions are highlighed in bold.

	Mean precision	Recall	F-Measure	MAP	GMAP
Documents	0.2124	0.1450	0.1384	0.0903	0.0005
Concepts	**0.4572**	0.391	**0.3848**	0.297	0.0634
RDF triples	0.0455	0.001	0.0021	0.001	0.0000
Snippets	0.0655	0.038	0.0409	0.024	0.0001

3.3 Task 2b – Phase B

Table 5 shows our official results[2] for all the five batches, for three types of questions: Yes/No, Factoid, and List. When considering the official measures – accuracy for Yes/No type questions; mean reciprocal rank (MRR) for Factoid type questions; and mean F-measure for List type questions – we achieved consistently better results than the two BioASQ baseline approaches. In addition, we obtained the highest results for the Yes/No-type questions in Batches 1 & 5, and for the Factoid-type questions in Batches 3 & 5 (highlighted in bold in Table 5).

[1] We did not submit results for Batch 2 – Phase A

[2] Official results for the Summary-type questions are not available in the case of "exact" answers at the time of writing. And no results have been released in the case of "ideal" answers for all questions.

Table 5. Official results of our submssions for the Phase B test sets in the case of "exact" answers. Our best results among all submissions are highlighed in bold.

Batch	Yes/No	Factoid			List		
	Accuracy	StrictAcc.	Lenient Acc.	MRR	Mean precision	Recall	F-Measure
Batch1	**0.9375**	0.1852	0.1852	0.1852	0.0618	0.0929	0.0723
Bacth2	0.8214	–	–	–	0.1596	0.2057	0.1618
Batch3	0.8333	**0.0417**	**0.1250**	**0.0833**	0.1195	0.1780	0.1373
Bacth4	0.8750	0.0938	0.1250	0.1042	–	–	–
Batch5	**1.0000**	**0.1379**	**0.1724**	**0.1466**	–	–	–

Note that there are no official evaluation results for our submissions for the Factoid-type questions in Batch 2 and List-type questions in Batches 4 and 5. This is likely due to a data format issue in our submissions.

4 Discussion & Conclusion

In BioASQ 2014 challenge on automated MeSH indexing, our learning-to-rank based approach shows improved and competitive performance among all participating systems. Moreover, we demonstrate that our learning-to-rank method is a general and robust framework that allows systematic integration of results from other methods for improved performance. When we include predicted results from a knowledge-based approach (MTI) and a text classification method, we were able to achieve the highest recall results by both flat and hierarchical measures while still maintaining high precisions. For instance, compared to one of the baseline systems – MTI First Line –we were able to achieve a much higher recall (59% vs. 49%) with almost the same level of precision (62%) (See Table 2 for details).

Our best results for Task 2b are noted in the "exact" answers to the Factoid-type questions (see Table 5) where we used results of our previously developed named entity recognition (NER) tools. In fact, this approach appears to perform better than relying on the gold-standard concepts from Phase A based on our comparative analysis.

In conclusion, we participated in both tasks of the BioASQ 2014 challenge where we are ranked among one of the top teams for both tasks. In the future, we are interested in exploring the opportunities of our high-performing MeSH prediction methods in practical applications (e.g. support instant MeSH indexing) and the roles of our state-of-the-art automated entity recognition tools in question answering tasks.

Acknowledgements

We would like to thank the BioASQ 2014 task organizers as well as authors of the NLM's MTI system for providing the task and baseline data. We also thank Dr. Ritu

Khare for her help on proofreading the manuscript. This research is supported by the NIH Intramural Research Program, National Library of Medicine.

References

1. Kim, S., Kim, W., Wei, C.-H., Lu, Z., Wilbur, W.J.: Prioritizing PubMed articles for the Comparative Toxicogenomic Database utilizing semantic information. Database: The Journal of Biological Databases & Curation 2012, bas042 (2012)
2. Lu, Z., Kim, W., Wilbur, W.J.: Evaluating relevance ranking strategies for MEDLINE retrieval. Journal of the American Medical Informatics Association 16, 32-36 (2009)
3. Leaman, R., Khare, R., Lu, Z.: NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with DNorm. Proceedings of the CLEF 2013 Evaluation Labs and Workshop. CLEF, Valencia, Spain (2013)
4. Leaman, R., Wei, C.-H., Lu, Z.: NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem. BioCreative IV Challenge Evaluation Workshop vol. 2, pp. 34 (2013)
5. Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., Hsu, C.-N., Tsai, R.T., Dai, H.-J., Okazaki, N.: The gene normalization task in BioCreative III. BMC bioinformatics 12, S2 (2011)
6. Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M.: The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC bioinformatics 12, S3 (2011)
7. Mao, Y., Van Auken, K., Li, D., Arighi, C.N., Lu, Z.: The gene ontology task at biocreative IV. Proceedings of the Fourth Biocreative Challenge Evaluation Workshop, vol. 1, pp. 119-127. BioCreative, Bethesda, Maryland (2013)
8. Arighi, C.N., Wu, C.H., Cohen, K.B., Hirschman, L., Krallinger, M., Valencia, A., Lu, Z., Wilbur, J.W., Wiegers, T.C.: BioCreative-IV virtual issue. Database 2014, bau039 (2014)
9. Wu, C.H., Arighi, C.N., Cohen, K.B., Hirschman, L., Krallinger, M., Lu, Z., Mattingly, C., Valencia, A., Wiegers, T.C., Wilbur, W.J.: BioCreative-2012 virtual issue. Database: The Journal of Biological Databases & Curation 2012, bas049 (2012)
10. Partalas, I., Gaussier, E., Ngomo, A.-C.N.: Results of the First BioASQ Workshop. Proceedings of the first Workshop on BioASQ, vol. 1094. BioASQ@CLEF, Valencia, Spain (2013)
11. Lu, Z., Kim, W., Wilbur, W.J.: Evaluation of query expansion using MeSH in PubMed. Information retrieval 12, 69-80 (2009)
12. Névéol, A., Doğan, R.I., Lu, Z.: Author keywords in biomedical journal articles. Proceedings of the American Medical Informatics Association Symposium, vol. 2010, pp. 537. Washington, D.C. (2010)
13. Doğan, R.I., Lu, Z.: Click-words: learning to predict document keywords from a user perspective. Bioinformatics 26, 2767-2775 (2010)
14. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the gene ontology. Nucleic acids research 33, W783-W786 (2005)

15. Huang, M., Lu, Z.: Learning to annotate scientific publications. Proceedings of the 23rd International Conference on Computational Linguistics, pp. 463-471. Association for Computational Linguistics, Beijing, China (2010)
16. Huang, M., Névéol, A., Lu, Z.: Recommending MeSH terms for annotating biomedical articles. Journal of the American Medical Informatics Association 18, 660-667 (2011)
17. Kim, W., Aronson, A.R., Wilbur, W.J.: Automatic MeSH term assignment and quality assessment. Proceedings of the American Medical Informatics Association Symposium, pp. 319, Washington, D.C. (2001)
18. Mork, J.G., Yepes, A.J.J., Aronson, A.R.: The NLM Medical Text Indexer System for Indexing Biomedical Literature. Proceedings of the first Workshop on BioASQ, vol. 1094. BioASQ@CLEF, Valencia, Spain (2013)
19. Névéol, A., Shooshan, S.E., Humphrey, S.M., Mork, J.G., Aronson, A.R.: A recent advance in the automatic indexing of the biomedical literature. Journal of biomedical informatics 42, 814-823 (2009)
20. Mao, Y., Lu, Z.: NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. Technical report (2013)
21. Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.P.: Large-Scale Semantic Indexing of Biomedical Publications. Proceedings of the first Workshop on BioASQ, vol. 1094. BioASQ@CLEF, Valencia, Spain (2013)
22. Wei, C.-H., Kao, H.-Y.: Cross-species gene normalization by species inference. BMC bioinformatics 12, S5 (2011)
23. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17, 229-236 (2010)
24. Wei, C.-H., Harris, B.R., Li, D., Berardini, T.Z., Huala, E., Kao, H.-Y., Lu, Z.: Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. Database: The Journal of Biological Databases & Curation 2012, bas041 (2012)
25. Wei, C.-H., Kao, H.-Y., Lu, Z.: PubTator: A PubMed-like interactive curation system for document triage and literature curation. proceedings of BioCreative 2012 workshop, pp. 145-150. BioCreative, Washington D.C. (2012)
26. Wei, C.-H., Kao, H.-Y., Lu, Z.: PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research 41, W518-W522 (2013)
27. Leaman, R., Doğan, R.I., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29, 2909-2917 (2013)
28. Wei, C.-H., Kao, H.-Y., Lu, Z.: SR4GN: a species recognition software tool for gene normalization. Plos one 7, e38460 (2012)
29. Wei, C.-H., Harris, B.R., Kao, H.-Y., Lu, Z.: tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 29, 1433-1439 (2013)
30. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. CoRR abs/1306.6802, (2013)