MeSHLabeler and DeepMeSH: Recent Progress in Large-scale MeSH Indexing

Shengwen Peng[1,2], Hiroshi Mamitsuka[3,4], Shanfeng Zhu*[1,2,5]

[1]School of Computer Science, Fudan University, Shanghai 200433, China.

[2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China.

[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan.

[4]Department of Computer Science, Aalto University, Espoo 02150 Finland.

[5]Center for Computational System Biology, Fudan University, Shanghai 200433, China

Email:

Shanfeng Zhu- zhusf@fudan.edu.cn;

*Corresponding author

MeSHLabeler and DeepMeSH: Recent Progress in Large-scale MeSH Indexing

Shengwen Peng[1,2], Hiroshi Mamitsuka[3,4], Shanfeng Zhu*[1,2,5]

[1]School of Computer Science, Fudan University, Shanghai 200433, China.

[2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China.

[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan.

[4]Department of Computer Science, Aalto University, Espoo 02150 Finland.

[5]Center for Computational System Biology, Fudan University, Shanghai 200433, China

Abstract

The U.S. National Library of Medicine (NLM) uses the Medical Subject Headings (MeSH) (See Note 1) to index almost all 24 million citations in MEDLINE, which greatly facilitates the application of biomedical information retrieval and text mining. Large-scale automatic MeSH indexing has two challenging aspects: the MeSH side and citation side. For the MeSH side, each citation is annotated by only 12 (on average) out of all 28000 MeSH terms. For the citation side, all existing methods, including Medical Text Indexer (MTI) by NLM, deal with text by bag-of-words, which cannot capture semantic and context-dependent information well. To solve these two challenges, we developed the MeSHLabeler and DeepMeSH. By utilizing "Learning to Rank" (LTR) framework, MeSHLabeler integrates multiple types of information to solve the challenge in the MeSH side, while DeepMeSH integrates deep semantic representation to solve the challenge in the citation side. MeSHLabeler achieved the first place in both BioASQ2 and

BioASQ3, and DeepMeSH achieved the first place in both BioASQ4 and BioASQ5 challenges. DeepMeSH is available at http://datamining-iip.fudan.edu.cn/deepmesh

Keywords: MeSH Indexing, Text Categorization, Multi-label Classification, Medical Subject Headings, MEDLINE, Machine Learning

1. Introduction

MEDLINE (See Note 2) is the largest biomedical literature database in the world, which contains more than 24 million citations. MeSH terms are used to index almost all MEDLINE citations [1], which is crucial in biomedical text mining and information retrieval [2-8]. The NLM annotators who are responsible for annotating the MeSHs need to review the full text of a citation, which costs lots of time and money. For the year 2016, there were 869,666 new citations in MEDLINE (See Note 3), and the average cost per citation was about $9.4 [9]. As of April 2016, there were 127 staff members in NLM who are responsible for annotating the most relevant MeSH terms to the MEDLINE citations [10]. As time goes on, the rapid increase of the MEDLINE citation poses great challenges for manual annotations. A fast and accurate automated MeSH indexing system is imperative to improve the indexing efficiency and reduce the cost.

NLM has developed an automated MeSH indexing system, MTI [11-13], to facilitate the annotation of MeSH. MTI mainly consists of two parts: MMI (MetaMap Indexing) [14] and PRC (PubMed Related Citations) [15]. MMI extracts the biomedical concept from the title and abstract, then maps it to the corresponding MeSH. Moreover, PRC tries to use the improved K-nearest Neighbor (KNN) algorithm to find the most similar MEDLINE citations, and then extracts MeSHs from these similar citations. The results of PRC and MMI were combined into a preliminary recommendation. After some processing (for example, the application of the index rules), MTI generate the final MeSH recommended list to the NLM annotators.

Large-scale MeSH indexing mainly has two aspects of challenges from the MeSH side and the citation side, respectively. In the MeSH side, the difference between the distribution of different MeSHs is particularly large. For example, among all 28,000 MeSH terms, the most common MeSH, "Humans", appears more than 8 million times in the MEDLINE, while a rare MeSH, such as "Pandanaceae", only appears 31 times. In addition, the number of MeSHs annotated for each citation varies greatly, which might be less than five MeSHs, or more than 30 MeSHs. In addition, in the side of citations, the “Bag of Words” method cannot effectively capture the complex semantics of biomedical documents because of the large number of concepts and abbreviations in biomedical literature. In many cases, similar concepts can be represented by different words, and the same words can express a completely different meaning from the context.

In order to promote the development of semantic indexing and automatic question answering systems in the biomedical field, BioASQ [16-18], a challenge on large-scale biomedical semantic indexing and question answering, held an international competition from 2013 to 2017. There have been many effective systems that have emerged through the platform, such as MetaLabeler [19], and MeSH Now [20]. To improve the performance of automatic MeSH indexing system, we developed two system: MeSHLabeler [21] and DeepMeSH [22], which solve the challenges in the MeSH side and citation side, respectively. MeSHLabeler use “Learning to Rank” framework to incorporates multiple evidences to rank the MeSHs while DeepMeSH integrates a new semantic representation to represent citations. MeSHLabeler achieved the first place in both BioASQ2 and BioASQ3, and DeepMeSH achieved the first place in both BioASQ4 and BioASQ5 challenges [23].

2. Materials

We use 2016 MeSH, containing 27,883 unique MeSH terms. Most of training data come from 2016 MEDLINE/PubMed baseline database downloaded from the NCBI website. Another part of data is downloaded from BioASQ 2015 challenge Task 3a, with 49,774 indexed. The text of all these citations only contain abstract, article title and journal title. DeepMeSH consists of two components, MeSHRanker and MeSHNumber. Given a target citation, MeSHRanker returns a ranked list of candidate MeSH terms, while MeSHNumber predicts the number of associated MeSH terms. For the 49,774 citations from BioASQ 2015, we randomly assign them to three sets: MeSHRanker training set (with 23,774 citations), MeSHNumber training set (with 20,000 citations) and local test set (with 6,000 citations).

Our system was mainly written by C++. It also used many open source tools to implement the whole flow.

1) BioTokenizer was used to tokenize and stem raw text.
2) LIBLINEAR was used to implement Logistic regression and Linear SVM.
3) XGBoost was used to implement learning to rank framework.

Our server has 4 Intel XEON E5-4650 2.7GHzs CPUs and 128G memory. It costs around 7 days to train 27,000 binary classifiers with logistic regression or linear svm. Predicting 10,000 citations costs around 3 hours.

3. Methods

DeepMeSH is the ‘state of the art’ of the MeSH indexing system. It improves the MeSH indexing accuracy by incorporating deep semantic representation to MeSHLabeler. The deep semantic representation, D2V-TFIDF, combines the advantages of the D2V (Document Vector) and TFIDF (Term Frequency with Inverse Document Frequency). According to our experiments, D2V-TFIDF represents citation texts better than both D2V and TFIDF. It is more powerful to find similar citations, so we use this representation to solve the challenge on the citation side.

MeSHLabeler is the last generation of MeSH indexing system, which uses the 'learning to rank' framework to incorporates multiple evidence to solve the challenge on the MeSH side. It has two components: MeSHRanker and MeSHNumber. MeSHRanker is used to rank the candidate MeSH terms for each target citation. On the other hand, MeSHNumber is used to predict the number of associated MeSH terms for the target citation. MeSHRanker incorporates five different types of evidence to rank the MeSHs, which includes global evidence, local evidence, MeSH dependency, pattern matching and MTI.

- Global Evidence: We train a binary classifier for each MeSH with the entire MEDLINE. Since each MeSH is trained independently, the scores returned by the different classifiers are theoretically incomparable. MeSHLabeler proposed a normalized method to deal with the score comparisons between different models, which significantly improves the prediction accuracy. Because each MeSH binary classifier is trained with the entire MEDLINE, we call this part of the evidence as global evidence.
- Local Evidence: For a target citation, we can score the candidate MeSHs through counting the MeSHs indexed by its similar citations.
- MeSH Dependency: It is a unique feature of MeSHLabeler that effectively considers the relevance of the MeSH-MeSH terms. For

infrequent MeSH terms, this information can effectively improve the accuracy of labeling. Since the number of MeSH-MeSH combinations was very large, none of the previous studies considered MeSH dependency.

- Patten Matching: We directly use the string matching method to find the MeSHs or their synonyms in the title or abstract.
- MTI: MTI considers not only pattern matching and local evidence, but also the index rules with domain knowledge. We integrate the results from MTI.

4. Usage

The input interface is shown in Figure 1.

1) Select a file. This is to upload the citations for MeSH indexing. Two file extensions, ".txt" and ".json", are supported. If the file extension is ".txt", the file must contain the PubMed IDs (pmid) of all target citations, and each line contains a pmid. If the file extension is ".json", the file should contain all raw texts of target citations. The text contains abstract and title. Note that each submit file must contain at least 100 instances. Sample input was shown as figure 2.
2) Input an email address. Input your email address to receive the prediction result. The email address will be used to receive a process id, prediction result or some information if any error occurs.
3) Upload the file. Click the submit button, the file will be uploaded. Once uploaded successfully, you will receive a process id in your email. As shown in figure 3, you can use the process id to check the prediction status.

For each successful submission. we will output a json file. For an unindexed pmid, we output the MeSHs recommended by DeepMeSH, and the answer_type has a value of "predicted". For an indexed pmid, we output the MeSHs indexed by PubMed, and the answer_type has a value of "annotated". If a pmid cannot be found in PubMed, the result will be empty, and the answer_type has a value of "not_found". Note that if the input file is a ".json" file which contains only texts, the output pmid is the index of the text in the file (starts from 0).

A sample output is shown as follows:

```
{"documents": [
{"labels":["D006801","D055815"],"pmid":24639323,"answer_type":"predicted"},
{"labels":["D000293"],"pmid":24687846,"answer_type":"predicted"},
{"labels":["D005260","D058006"],"pmid":27059885,"answer_type":"annotated"},
{"labels":[],"pmid":32131231, "answer_type":"not_found"}]}
```

4. Notes

1. https://www.nlm.nih.gov/pubs/factsheets/mesh.html

2. https://www.nlm.nih.gov/pubs/factsheets/medline.html

3. http://www.nlm.nih.gov/bsd/bsd_key.html

Acknowledgements

This work has been partially supported by National Natural Science Foundation of China (Grant Nos: 61572139), MEXT KAKENHI #16H02868 and FiDiPro by Tekes.

References

[1] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ (2004) The NLM Indexing Initiativeś Medical Text Indexer. Stud Health Technol Inform 107(Pt 1): 268-272

[2] Stokes N, Li Y, Cavedon L, Zobel J (2010) Exploring criteria for successful query expansion in the genomic domain. Information Retrieval 12: 17-50

[3] Lu Z, Kim W, Wilbur WJ(2010) Evaluation of query expansion using MeSH in PubMed. Information Retrieval 12 : 69-80

[4] Zhu S, Takigawa I, Zeng J, Mamitsuka H (2009) Field independent probabilistic model for clustering multi-field documents. Inf. Process. Manage. 45(5): 555-570

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

[6] Gu J, Feng W, Zeng J, Mamitsuka H, Zhu S (2013) Efficient Semisupervised MEDLINE Document Clustering With MeSH-Semantic and Global-Content Constraints. IEEE Trans. Cybernetics 43(4): 1265-1276

[7] Zhou J, Shui Y, Peng S, Li X, Mamitsuka H, Zhu S (2015) MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents. J. Bioinformatics and Computational Biology 13(6)

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

[9] Mork JG, Jimeno-Yepes A, Aronson AR (2013) The NLM Medical Text Indexer System for Indexing Biomedical Literature. *BioASQ@ CLEF*

[10]Demner-Fushman D, Mork JG (2016) A Report to the Board of Scientific Counselors, April 2016

[11]Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ (2004) The NLM Indexing Initiativeś Medical Text Indexer. Stud Health Technol Inform 107(Pt 1): 268-272

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

[13]Nelson SJ, Schopen M, Savage AG, Schulman JL, Arluk N (2004) The MeSH translation maintenance system: structure, interface design, and implementation. Medinfo 11: 67-69

[14]Aronson AR, Lang FM (2004) An overview of MetaMap: historical perspective and recent advances. J Am Med Infor Assoc 17: 229-236

[15]Lin J, Wilbur WJ (2007) PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 8: 423.

[16]Partalas I, Gaussier É, Ngomo ACN et al. (2013) Results of the First BioASQ Workshop. BioASQ@ CLEF

[17]Tsatsaronis G et al. (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics, 16: 138

[18]Balikas G, Partalas I, Ngomo AN, Krithara A, Paliouras G (2014) Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. CLEF (Working Notes): 1181-1193

[19]Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas IP (2013) Large-Scale Semantic Indexing of Biomedical Publications. BioASQ@ CLEF

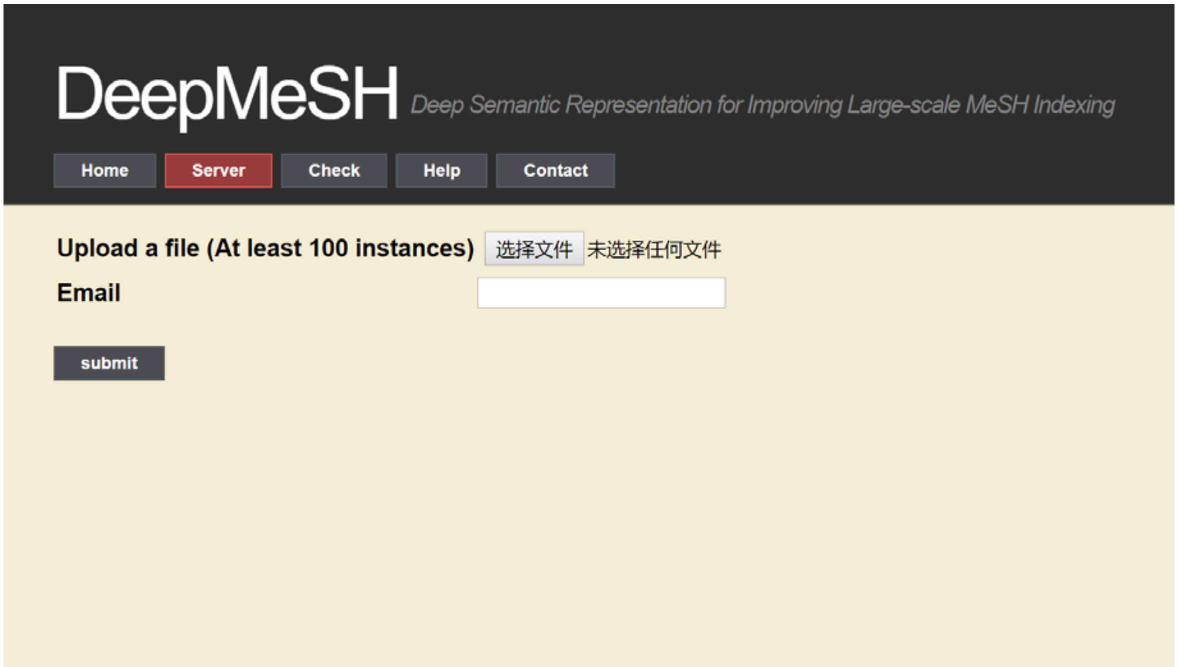
[20]Mao Y, Lu Z (2013) NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. BioASQ@ CLEF

[21]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 12: i339-347

[22]Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S (2016) DeepMeSH: deep semantic

representation for improving large-scale MeSH indexing. *Bioinformatics* 32(12): i70–i79

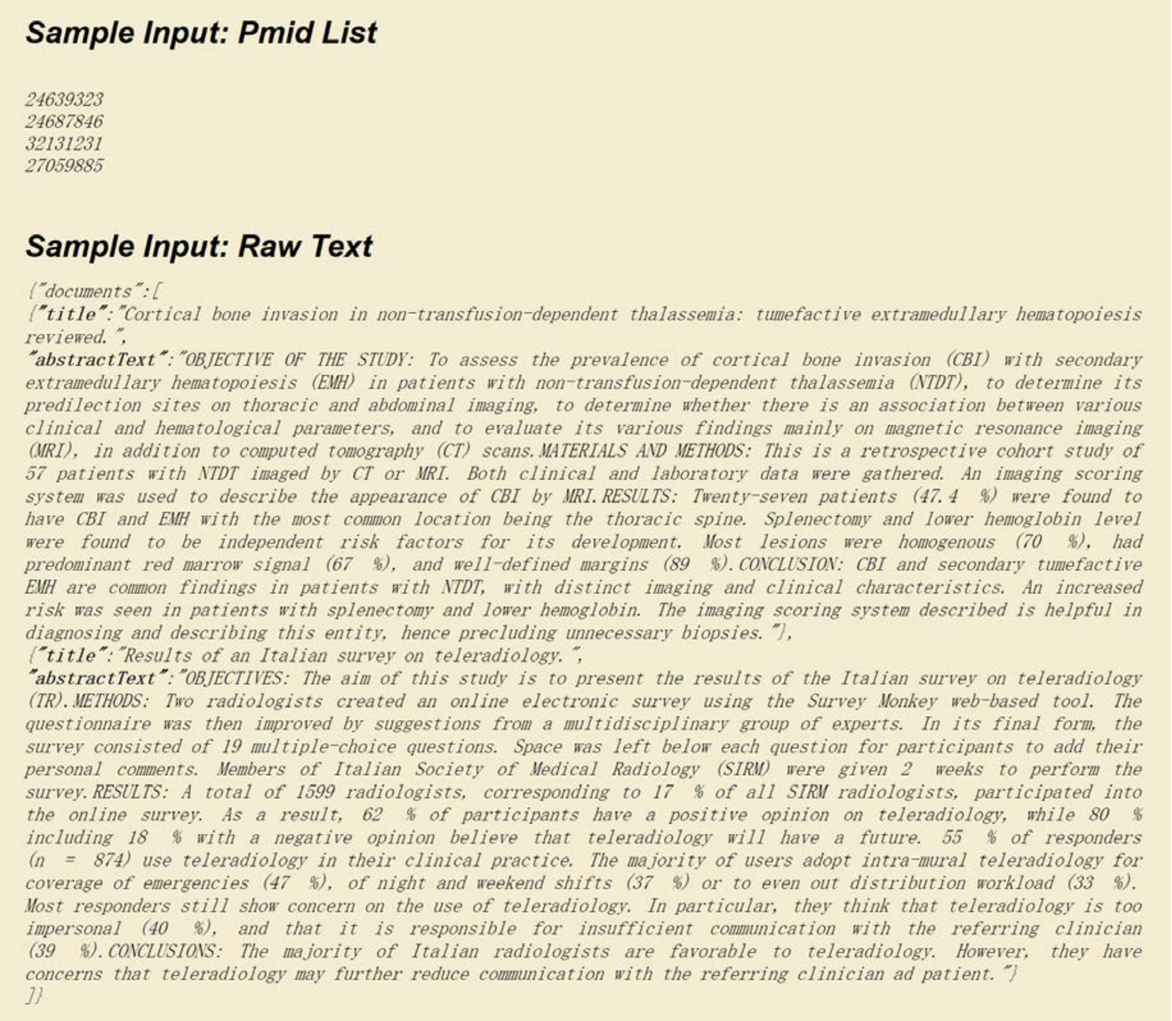
[23] Peng S, You R, Xie Z, Wang B, Zhang Y, Zhu S (2015) The Fudan Participation in the 2015 BioASQ Challenge: Large-scale Biomedical Semantic Indexing and Question Answering. CLEF (Working Notes)


DeepMeSH
Deep Semantic Representation for Improving Large-scale MeSH Indexing
Home
Server
Check
Help
Contact
选择文件
submit

Upload a file (At least 100 instances) 未选择任何文件

Email

Figure 1: The input interface of DeepMeSH



Sample Input: Pmid List

```
24639323
24687846
32131231
27059885
```

Sample Input: Raw Text

```
{"documents":[
{"title":"Cortical bone invasion in non-transfusion-dependent thalassemia: tumefactive extramedullary hematopoiesis
reviewed.",
"abstractText":"OBJECTIVE OF THE STUDY: To assess the prevalence of cortical bone invasion (CBI) with secondary
extramedullary hematopoiesis (EMH) in patients with non-transfusion-dependent thalassemia (NTDT), to determine its
predilection sites on thoracic and abdominal imaging, to determine whether there is an association between various
clinical and hematological parameters, and to evaluate its various findings mainly on magnetic resonance imaging
(MRI), in addition to computed tomography (CT) scans.MATERIALS AND METHODS: This is a retrospective cohort study of
57 patients with NTDT imaged by CT or MRI. Both clinical and laboratory data were gathered. An imaging scoring
system was used to describe the appearance of CBI by MRI.RESULTS: Twenty-seven patients (47.4 %) were found to
have CBI and EMH with the most common location being the thoracic spine. Splenectomy and lower hemoglobin level
were found to be independent risk factors for its development. Most lesions were homogenous (70 %), had
predominant red marrow signal (67 %), and well-defined margins (89 %).CONCLUSION: CBI and secondary tumefactive
EMH are common findings in patients with NTDT, with distinct imaging and clinical characteristics. An increased
risk was seen in patients with splenectomy and lower hemoglobin. The imaging scoring system described is helpful in
diagnosing and describing this entity, hence precluding unnecessary biopsies."},
{"title":"Results of an Italian survey on teleradiology.",
"abstractText":"OBJECTIVES: The aim of this study is to present the results of the Italian survey on teleradiology
(TR).METHODS: Two radiologists created an online electronic survey using the Survey Monkey web-based tool. The
questionnaire was then improved by suggestions from a multidisciplinary group of experts. In its final form, the
survey consisted of 19 multiple-choice questions. Space was left below each question for participants to add their
personal comments. Members of Italian Society of Medical Radiology (SIRM) were given 2 weeks to perform the
survey.RESULTS: A total of 1599 radiologists, corresponding to 17 % of all SIRM radiologists, participated into
the online survey. As a result, 62 % of participants have a positive opinion on teleradiology, while 80 %
including 18 % with a negative opinion believe that teleradiology will have a future. 55 % of responders
(n = 874) use teleradiology in their clinical practice. The majority of users adopt intra-mural teleradiology for
coverage of emergencies (47 %), of night and weekend shifts (37 %) or to even out distribution workload (33 %).
Most responders still show concern on the use of teleradiology. In particular, they think that teleradiology is too
impersonal (40 %), and that it is responsible for insufficient communication with the referring clinician
(39 %).CONCLUSIONS: The majority of Italian radiologists are favorable to teleradiology. However, they have
concerns that teleradiology may further reduce communication with the referring clinician ad patient."}
]}
```

Figure 2: A sample input of DeepMeSH

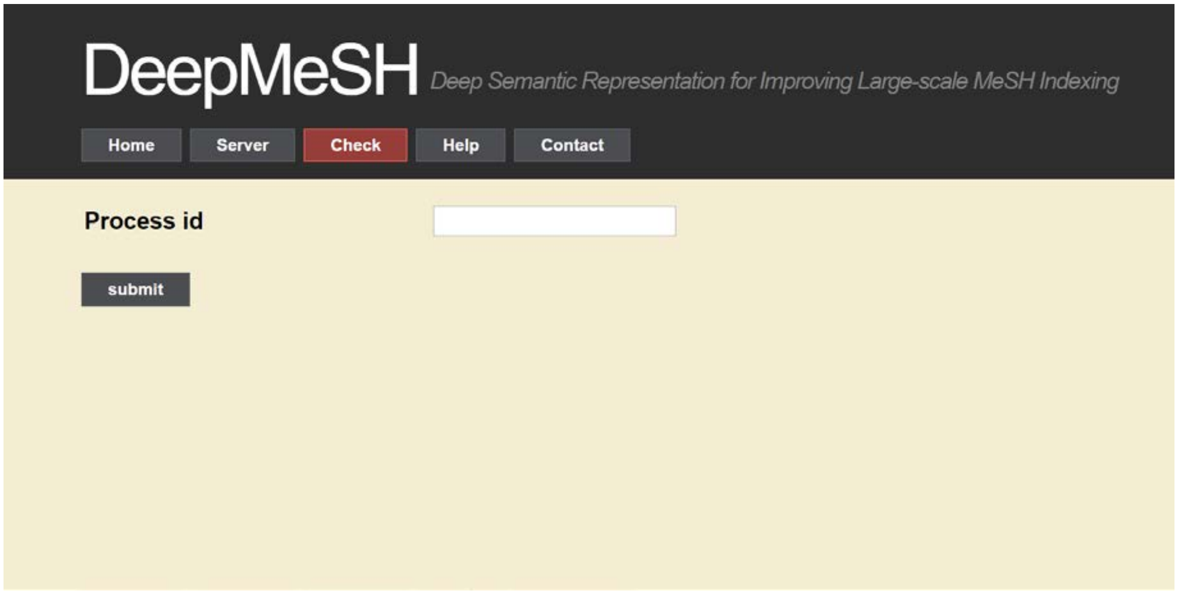

DeepMeSH
Deep Semantic Representation for Improving Large-scale MeSH Indexing
Home
Server
Check
Help
Contact
Process id
submit

Figure 3: The process id check interface of DeepMeSH