Journal of Biomedical Informatics 93 (2019) 103143

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Multi-label biomedical question classification for lexical answer type prediction

Muhammad Wasim[a,b,d,*], Muhammad Nabeel Asim[b,c], Muhammad Usman Ghani Khan[a,b], Waqar Mahmood[b]

[a] Dept. of Computer Science & Engineering, UET, Lahore, Pakistan
[b] Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan
[c] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[d] Department of Computer Science, UMT Lahore, Sialkot Campus, Pakistan

ARTICLE INFO

*Keywords:*
Biomedical question classification
Lexical answer type prediction
Biomedical LAT corpus
Multi-label classification

ABSTRACT

Question classification is considered one of the most significant phases of a typical Question Answering (QA) system. It assigns certain answer types to each question which leads to narrow down the search space of possible answers for factoid and list type questions. The process of assigning certain answer types to each question is also known as Lexical Answer Type (LAT) Prediction. Although much work has been done to enhance the performance of question classification into coarse and fine classes in diverse domains, it is still considered a challenging task in the biomedical field. The difficulty in biomedical question classification stems from the fact that one question might have more than one label or expected answer types associated with it (also, referred to as a multi-label classification). In the biomedical domain, only preliminary work is done to classify multi-label questions by transforming them into a single label through copy transformation technique. In this paper, we have generated a multi-labeled corpus (MLBioMedLAT) by exploring the process of Open Advancement of Question Answering (OAQA) system for the task of biomedical question classification. We use 780 biomedical questions from BioASQ challenge and assign them appropriate labels. To annotate these labels, we use the answers for each question and assign the question semantic type labels by leveraging an existing corpus and utilizing OAQA system. The paper introduces a data transformation approach namely *Label Power Set* with logistic regression (LPLR) for the task of multi-label biomedical question classification and compares its performance with Structured SVM (SSVM), Restricted Boltzmann Machine (RBM), and copy transformation based logistic regression (CLR) (previously used for a similar task in the OAQA system). To evaluate the integrity of the introduced data transformation technique, we use three prominent evaluation measures namely *Micro* $F_1$, *Accuracy*, and *Hamming Loss*. Regarding *Micro*$F_1$, our introduced technique coupled with a new feature set surpasses CLR, SSVM, and RBM with a margin of 7%, 8%, and 22% respectively.

## 1. Introduction

The recent explosion of scientific literature, specifically in the biomedical domain, has raised question over the intelligence and capabilities of search engines [1]. Such limitations demonstrate a desperate need of Question Answering (QA) systems capable of providing precise information instead of documents' list[2]. The study to develop question answering systems dates back to 1960s [3–5] and the current resurgence in QA systems is motivated by the continuous growth of the massive amount of scientific and non-scientific literature [6]. Biomedical experts are particularly interested in finding the precise information to make smooth research progress and avoid repeating the same experiments conducted by other researchers [7,8]. A typical QA system comprises three modules namely question processing, candidate retrieval, and answer processing [2]. Fig. 1 depicts the pipeline of a standard question answering system. In this paper, we focus on the question classification step which is part of the question processing phase.

Biomedical question classification can be performed either to determine the *question type* [9] or *lexical answer type (LAT)* [10–12] of a question. Question type classification is used to determine the question type such as *factoid*, *list*, *yes/no* or *summary*. Such information facilitates in the use of an appropriate strategy for answering a question [9]. Question classification for *lexical answer type* prediction is concerned with determining the expected biomedical entity that the question is

* Corresponding author at: Department of Computer Science, UMT Lahore, Sialkot Campus, Pakistan.
*E-mail address:* wasim@kics.edu.pk (M. Wasim).

https://doi.org/10.1016/j.jbi.2019.103143
Received 2 August 2018; Received in revised form 28 February 2019; Accepted 3 March 2019
Available online 12 March 2019
1532-0464/ 

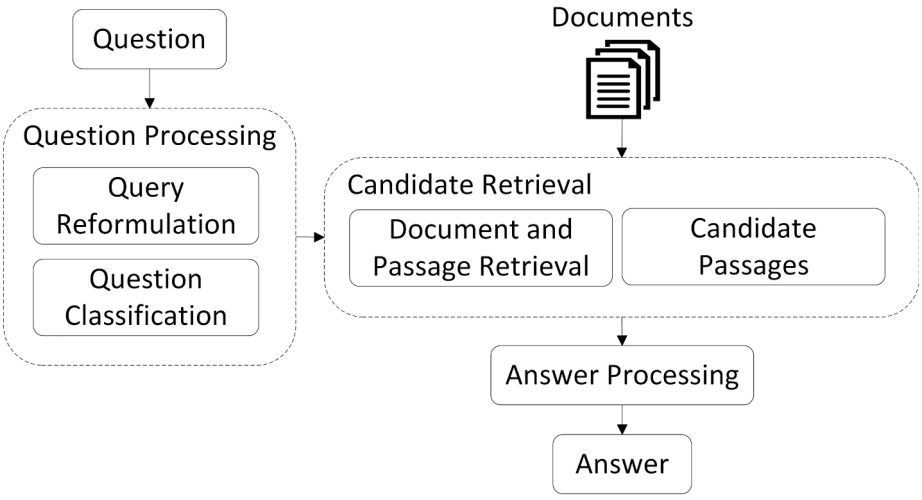


**Fig. 1.** Pipeline of a typical question answering system.

seeking for. For example, for a question, *When was empagliflozin FDA approved?*, the task of lexical answer type prediction is to assign a class *temporal concept* to the question. Question LAT prediction aims to reduce search space for selecting appropriate candidate answers in the answer processing stage. The improvements in LAT prediction may lead to a positive impact on the answer processing stage as the right answer types will be narrowed down thus leading to the increased QA system performance.

Question classification for lexical answer type prediction can be categorized as: *rule based*, *learning based* and *hybrid* approaches [13,14]. Rule-based techniques match the specified questions with the manually crafted or defined rules. Rule-based techniques do not perform well on diverse datasets as it requires numerous rules to classify the specified questions successfully. Learning-based approaches classify the specified questions by using a pipeline of various tasks: extraction of features from the questions, training of the classifier and prediction of entity type or class labels by exploiting a trained classifier. Learning based techniques can be further divided into supervised and semi-supervised techniques. Finally, hybrid approaches employ both learning and rule-based methodologies in quest of question classification. In different domains, researchers worked on all three (rule-based, learning based, hybrid) mentioned techniques for question classification to improve the performance of question answering system [15,13,14,16]. However, question classification in biomedical domain still requires substantial research efforts.

Question classification for determining the lexical answer type or expected *semantic type* of a question has gained good accuracy in open domain [17,14] but it is still challenging in the biomedical domain as a typical biomedical question may tend to have more than one answer types associated with it. Let us take a simple biomedical question as an example: *Which are the enzymes involved in the addition of 7-methylguanosine in mRNA?*. For this question, the lexical answer type might be *Enzyme*, *Chemical*, *Gene*, *Biologically Active Substance*, or *Amino Acid, Peptide, or Protein*. To further highlight the need of multi-label question classification, another question from BioASQ corpus - *Which are the supplemental antioxidant in athletes?* – has 19 possible answers[1]. Let us take three answers as an example: *Creatine*, *coenzyme Q(10)*, and *Resveratrol*. The UMLS terminology service (UTS) returns six semantic types for these answers as shown in Table 1. These examples demonstrate that assuming a single expected answer type for the biomedical question is doomed to failure. The problem of predicting multiple labels for a given biomedical question is a more specific case of a general task known as *multi-label learning*.

Multi-label learning is a paradigm of supervised learning in which a document belongs to more than one class [18]. Tsoumakas et al. [19] categorized it into two different classes: Label Ranking (LR) and Multi-label Classification (MLC). In label ranking, all possible labels for a given instance x are ranked based on the following function h: $X \times T \rightarrow W$ where $T = \{\alpha_1, \alpha_2, ..., \alpha_r\}$ is an output space of $r > 1$ labels and X is an $e$ dimensional feature space. Each subset of T is known as label-set. If $h(x, \alpha_1) > h(x, \alpha_2)$ then, $\alpha_1$ will be ranked higher than $\alpha_2$ for an input x. MLC is a particular type of multi-label classification characterized as a function $h: X \rightarrow 2^T$. Classifier gives relevant R set of labels for a particular instance x which is given as an input. Furthermore, Multi-label classification can be performed using two different group of methods: *data transformation* and *method adaptation* techniques [20]. Data transformation techniques are algorithm independent in that they transform the provided multi-label data into one or more single label classification task(s). Later, standard classification algorithms are employed to classify the data. In data transformation, we transform multi-label data into single label using different transformation techniques such as *Copy*, *Copy Weight*, *Binary Relevance*, and *Label Powerset*. Once transformed, we can use state of the art multi-class classification algorithms for classification. Researchers have used data transformation techniques in multi-label classification for music categorization, image classification, and text classification [20]. The method adaptation technique extends the existing multi-class classification algorithms to deal with the provided multi-label data directly. Such extensions already exist for decision trees [21], support vector machines [20], and neural networks [22]. In our experiments, we have used both data transformation and method adaptation techniques to improve the accuracy of multi-label question classification.

We build on top of the work previously performed by Neves et al. [12] and Yang et al. [23]. The existing corpus for biomedical question LATs provides single lexical answer type for 99% of the questions [12]. We show the limitation of assigning single LAT to biomedical questions and improve its quality by adding more expected answer types and adding new questions with LATs using the corpus generation process available in OAQA. Since OAQA is a complicated system, we explore the system components, generate multi-label question corpus, perform multi-label classification and provide the corpus online[2] for further experimentation. The motivation behind our study is to answer one fundamental question: *How multi-label biomedical question classification can be improved?*. To the best of our knowledge, only OAQA exploits the multi-label nature of biomedical LAT prediction using copy transformation technique, but they did not report the results of this specific

[1] http://participants-area.bioasq.org/Tasks/b/testData/

[2] https://github.com/wasimbhalli/Multi-label-Biomedical-QC-Corpus.

**Table 1**
Semantic Types for Three Answers of a Biomedical Question; multiple types separated by semi-colon and their identifiers are shown in parenthesis.

| Answer | Semantic types |
|---|---|
| Creatine | Amino Acid, Peptide, or Protein (T116); Biologically Active Substance (T123) |
| Coenzyme Q(10) | Organic Chemical (T109); Biologically Active Substance (T123) |
| Resveratrol | Organic Chemical (T109); Pharmacologic Substance (T121) |

step. We analyzed the copy transformation technique for biomedical question classification and found that it suffers from three main problems: (1) the system does not exploit the dependence between labels of a particular question, (2) the classifier suffers from decision boundary construction during training with copy transformation based technique and (3) the predicted number of labels is fixed (five).

We introduce label power set (LP) transformation technique with logistic regression (LPLR) to overcome the problems mentioned above and compare its performance with state of the art copy transformation based logistic regression (CLR) previously used in OAQA. To further evaluate the integrity of the proposed method, we compare its performance with other famous multi-label classification techniques (SSVM, RBM). Our experiments show that LPLR outperforms CLR, SSVM, and RBM with the margin of 4%, 5%, and 19% respectively. We also explore the effect of eight features used in OAQA and perform feature engineering which not only increases the performance of introduced technique but also improves the performance of other compared methods. Overall, our introduced data transformation technique along with feature engineering outperforms the copy transformation technique used for biomedical question classification in the OAQA system with a margin of 7%. Integration of our proposed methodology in the pipeline of the OAQA system also shows promising results for the complete QA system.

## 2. State of the art work

Although the work on question classification has been studied extensively in open domain [15,24,16,25,14], the work which directly tackles the challenge of multi-label question classification in biomedical domain is minimal. In the biomedical domain, researchers have worked on question type classification to classify questions into factoid, list, summary and yes/no classes automatically [9]. Moreover, the researchers in this area mostly adopted the work from the open domain and applied it to biomedical question classification. For example, researchers worked on manual lexical answer type annotation using headwords [12], rule-based answer type identification [11], and machine learning based question classification [26]. Yang et al. [23] did a unique work in this area by generating biomedical question classification corpus and performed multi-label classification which is also the focus of our study and need of the hour.

In the open domain, the work of Li and Roth [15,27] was one of the first attempts to use supervised machine learning for question classification. The researchers employed two-level taxonomy (6 coarse classes and 50 fine-grained classes) with multi-class classification. They manually created a list of semantic features and used very high dimensional feature space for classification. They compiled their dataset from existing open domain TREC dataset (later referred to as UIUC dataset) and applied the Winnow algorithm for classification [28]. Haung et al. [16], proposed another approach as useful as presented by Li and Roth with a more compact feature set such as wh-word, headword (syntactic and semantic), word grams and word shape. In particular, they presented headword feature and used two approaches (direct and indirect hypernyms) on a dataset to augment semantic features using WordNet. For the experimentation, they used SVM and maximum entropy on UIUC dataset which contained 5500 questions. Finally, Silva et al. [14] presented an effective use of rule-based classifier (standalone, as a feature in the machine learning approach). They used rule-based techniques for two purposes: pattern matching and headword extraction. In pattern matching, a question is directly mapped to a category if it follows a specific pattern. With this pattern matching, they obtained the precision of 99.9% in coarse-level classes and 98.0% in fine-level classes. Whereas in headword extraction, where headwords are extracted by using the rule-based parser and then mapped to a question category, they were able to get 86.4% in coarse classes and 78.5% in fine-level classes using 50 synset clusters. Moreover, they integrated the results of rule-based parser as a feature in the machine learning technique (SVM) and generated 95.0% in coarse-level classes and 90.8% in fine-level classes. They utilized a very compact feature set of 10,000 features as compared to previous work (200,000 features) [15] giving results with increased accuracy.

For the biomedical domain, Sarrouti et al. [9] presented a syntactic and rule-based technique (set of patterns) to classify biomedical questions into three classes: yes/no, factoid and summary. They used Stanford's POS Tagger to define the structure of question and MetaMap for biomedical named entity recognition (BNER) on the dataset of 1433 questions collected from BioASQ challenge. Regular expressions were used to classify all type of questions. To improve the classification task, they used WordNet to extend features with synonyms. Moreover, they utilized MetaMap service to map question terms to UMLS (biomedical vocabulary) to extract named entities. Overall, the system achieved an accuracy of 91.62%. The system attained lowest accuracy (69.21%) for "what" type question, and highest accuracy for *yes/no* type of questions. The study focused on question type instead of lexical answer types. Although such research can help in defining answer search strategy, for factoid and list type of questions, lexical answer type is required to find expected answer entities.

Mariana Neves and Milena Kraus [12] manually annotated 643 questions (BiomedLAT corpus) with headwords and UMLS semantic types based on the headwords and answers provided in the BioASQ training data. They assigned a single semantic type to most of the questions with only four questions with two semantic types. They performed annotations with 343 headwords and 53 distinct UMLS semantic types. In their initial experiments, they could assign a correct semantic type to only 184 questions (28.6%). The BiomedLAT corpus provides a single label for every question for 639 biomedical questions but, in reality, the gold standard BioASQ questions demonstrate the apparent need of more than one expected answer type for each question as discussed in Section 3. During the annotation process, the annotators did not consult any biomedical terminology source to add all possible entity types for given biomedical questions. Moreover, extending the manually annotated corpus requires considerable human effort.

Two biomedical question answering systems have employed single label based LAT predictions for biomedical questions [11,26]. Fudan system [11] defined regular expressions to extract the target semantic type, and so the system did not require any training corpus. These handcrafted regular expressions were able to identify few lexical answer types such as *gene* or *disease*. HPI [26] was another system which used machine learning to predict lexical answer type for a limited number of biomedical semantic types. They used headwords and lexical features to predict single expected semantic type using Support Vector Machine (SVM). The problem with their approach is that the number of

expected answer types was minimal as they only a single entity type as expected answer type.

Yang et al. developed their answer type prediction corpus automatically for their system - OAQA - which participated in BioASQ challenge for two consecutive years [29,23]. They used two services (UTS, tmtool) to lookup for the semantic types of all the answers and applied a threshold to select those semantic types which occurred more than 50% for a particular question. In addition to the UMLS semantic types, they defined three custom types: *choice*, *quantity* and *unlabeled*. The choice type of questions were those where the answer was present in the question. Quantity type of question expected a quantity. All the remaining questions which did not belong to UMLS semantic type, choice, or quantity type were assigned *unlabeled*. To train the logistic regression based model, they used copy transformation to train multi-label biomedical question and selected the top 5 lexical answer type. The problem with copy transformation for question classification is that multiple copies of the same question are created, and each copy is assigned different labels confusing the classification algorithm. They also did not report any results specific to lexical answer type prediction.

Although Neves et al. [12] developed a LAT corpus for biomedical questions, the corpus assigns a single label to 639 out of 643 questions. By manually inspecting the biomedical questions dataset provided by BioASQ, we found that one question usually has more than one answer types. We show the problem associated with existing dataset with the help of an example in Section 3.1. On the other hand, the corpus generation process of Yang et al. [23] generates too many expected answer types for individual questions even when they set its threshold to 50%. By taking advantage of both techniques, we developed a new corpus and analyzed eight features provided to the logistic regression based model after converting them to single-label using label power set based transformation technique. We describe the corpus generation process in the next section.

## 3. MLBioMedLAT: multi-label biomedical LAT corpus

This section briefly describes the limitation with BioMedLAT corpus prepared by Neves et al. [12] and details our process of corpus generation in which we leverage BioMedLAT corpus and extend it by a similar process previously used by Yang et al. [23].

### 3.1. Limitation in existing corpus

Neves et al. [12] developed a manually annotated corpus (BioMedLAT) of biomedical questions with their respective LATs in which only 4 out of 643 questions were bi-labeled, and rest of the questions were assigned a single label. Although they provided annotators with gold standard answers for each question, the annotators did not search the semantic type of answers in any biomedical lexicon such as unified medical language system (UMLS) to find the semantic types of all answers available for a particular question. Searching of semantic types of answers against a question reveals that biomedical questions expect more than one semantic type for each question. Three example questions from the BioMedLAT corpus are given below:

1. Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?
2. Which are the drugs utilized for the burning mouth syndrome?
3. List the main proteases used for sample digestion in proteomics.

Table 2 shows the semantic types assigned by the manual annotation process versus the semantic types identified by passing the exact answer from the gold corpus to the UMLS terminology service (UTS).

As shown in Table 2, both first and second question have been assigned the semantic type of *Clinical Drug* in the BioMedLAT corpus. UTS search for answer types of the available answers shows different answer semantic types returned for every question. For the second question, *Clinical Drug* did appear but only a single time and other types were more probable expected answer types for this question. For the third question, the semantic type *Enzyme* did match with the most likely type returned by UTS but another type, *Amino Acid, Peptide, or Protein* with the same probability of being the answer type is returned. The BioASQ dataset is prepared by experts keeping in view the information needs of biomedical professionals. The UMLS terminology service provides the semantic types for these answer entities which are again developed by biomedical experts. The search of gold standard answers from the UTS service reveals that the questions in the biomedical domain usually have more than one expected answer types. So, there is a need of multi-label corpus for biomedical question classification. We extend BioMedLAT dataset by adding more semantic types to existing questions and adding new questions from the 4th year of BioASQ challenge. We use the process similar to OAQA system [29] for multi-label biomedical corpus generation as described in the next section.

**Table 2**
Comparison of the Semantic Types assigned by BioMedLAT and UTS; the number in parenthesis represents the number of times a semantic type was found against multiple answers with each semantic type separated by a semicolon.

| Question no. | BioMedLAT semantic type | UTS semantic types |
|---|---|---|
| 1 | Clinical Drug | Organic Chemical (2); Pharmacologic Substance(2) |
| 2 | Clinical Drug | Pharmacologic Substance (2); Therapeutic or Preventive Procedure (2); Organic Chemical (2); Pharmacologic Substance (2); Clinical Drug (1) |
| 3 | Enzyme | Amino Acid, Peptide, or Protein(4); Enzyme (4); Pharmacologic Substance (2) |

### 3.2. The corpus generation process

In our experimentation, we use a similar process of corpus generation previously performed by Yang et al. [29] in OAQA. The OAQA system is very complicated as it comprises various modules such as question classification, document and passage retrieval, and answer selection. Yang et al. [29] used all these modules for the task of biomedical question answering. However, they neither elaborated the process of corpus generation for multi-label biomedical question classification nor discussed the results of this particular module. Therefore, the focus of our study is to explain and improve on the process of corpus generation and use it to extend manually curated BioMedLAT corpus.

To generate multi-label biomedical question classification corpus, we need questions and a classification scheme to annotate these questions. BioASQ is biomedical semantic indexing and question answering challenge initiated in 2013, and since then, it is being held every year. For our research, we utilize the training questions dataset available for BioASQ 5th year challenge[3] which includes all questions from last four years. It provides a total of 1799 questions that contain four types of questions: factoid, list, yes/no, and summarization. However, we only selected factoid, and list type of questions totaling 899 questions (factoid:486, list:413) as these type of questions require lexical answer types during the answer processing stage. Only a subset of these questions (780) was annotated automatically as the semantic type of 119 questions' answers could not be found.

For biomedical question answering, the classification scheme should be able to help in determining the semantic types of the expected

[3] http://participants-area.bioasq.org/Tasks/5b/trainingDataset/.

**Table 3**
Custom defined types with example question.

| Question type | Example question |
|---|---|
| Choice | Do archaeal genomes contain one or multiple origins of replication? |
| Quantity | What is the number of long non coding RNAs in the human genome? |

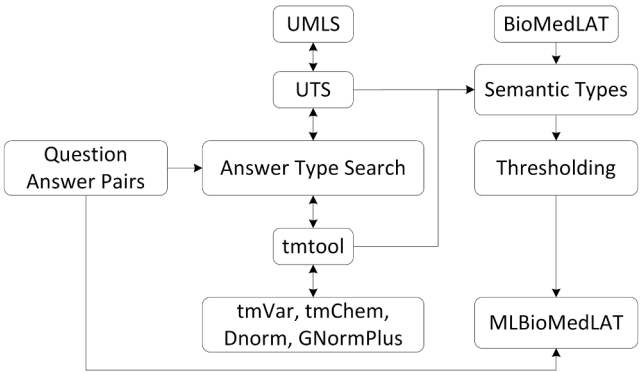


**Fig. 2.** Corpus generation process.

answers. As Metamap can be used in the later stages of a QA system to identify UMLS semantic types and tmtool is used to identify five different semantic types, we use the types available for both tools as our classification scheme (similar to OAQA). Overall, the total semantic types are 138, i.e. 133 from UMLS and five from tmtool. However, only a subset of these types (85) is used for classification as only these types are required to annotate the questions available in the corpus. Similar to OAQA, we add two additional question labels namely *choice* and *quantity* (we ignore the unlabeled questions). *Choice* and *Quantity* type of questions are identified using a set of self defined rules (Appendix A). The questions that contain tokens of *is/be*, *do*, and *or* are given the class of *choice*. On the other hand, *quantity* class is given to the questions that contain phrases such as *how many*, *how long* and *percentage*. Quantity type questions have some numeric value for their gold standard answers. Table 3 shows example questions for the custom defined types.

Fig. 2 shows the complete corpus generation process. The process starts by providing questions and their corresponding answers to the corpus generation system. The corpus generation process searches all the answers from the UMLS terminology service (UTS) and tmtool. To differentiate UTS and tmtool labels, the labels are prefixed with *umls* and *tmtool* respectively. The system utilizes two types of matching to return semantic types from UTS service: *exact* and *word*. In *exact match* approach, the system finds semantic types of the answers that match exactly. However, in *word* match approach, if any word of the complete answer phrase returns matches with a semantic type, it is also considered. For our corpus, we found 1811 exact and 397 word matches. We also incorporated BioMedLAT corpus by including all the lexical answer types available in it. We observe that for the questions where both BioMedLAT and MLBioMedLAT had a semantic type overlap, on average for every question, the semantic type occurred 65% of the times in automatically generated corpus. Therefore, we selected a threshold of 65%, so we selected all the semantic types which appeared more than 65% of the times for a particular question as the expected lexical answer type for that question. We excluded all the questions which could not be assigned any type during this process. The corpus is available online[4] for further experimentation and reproduction of results.

Table 4 shows 85 question classes and the number of questions belonging to each class with custom defined categories shown in bold. The table demonstrates that the generated dataset is highly skewed as 39 classes have three or fewer questions annotated with that particular class. Only six categories contain 50 or more questions. The dataset contains many questions which have more than one label assigned to them. For example, “Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?” is a tri-label question (UMLS:orch, umls:clnd, umls:phsu) whereas “Which are the plant DNA (cytosine-5) methyltransferase families?” is a penta-label question (umls:dsyn, umls:gngm, umls:bacs, umls:phsu, umls:aapp). The complete corpus generation algorithm is available in Appendix A.

## 4. Materials and methods

This section presents the proposed methodology of lexical answer type prediction for biomedical questions in multi-label learning paradigm. The complete process is depicted in Fig. 3. As the figure suggests, the process starts with preprocessing and feature extraction. To perform the multi-label classification, we use four different models. Two models are transformation based which first transform the data from multi-label to single-label using copy and label power set transformation techniques and, later, use the logistic regression classifier. We name these methods as copy based logistic regression (CLR) and label power set based logistic regression (LPLR). The third model, Structured SVM (SSVM), is based on the adaptation of SVM to deal with multi-label data directly. Finally, the fourth model is based on deep learning which uses a Restricted Boltzmann Machine (RBM) with logistic regression for each neuron at the output layer. Next subsections present a detailed description of these steps.

### 4.1. Preprocessing

We perform lemmatization using ClearNLP.[5] Moreover, each question is preprocessed using POS tagger and parser using ClearNLP Bioinformatics model.[6]

### 4.2. Feature extraction

Questions are a form of textual data which does not provide any features to discriminate questions using all possible classes and so we need feature extraction. This study extracts and analyzes eight feature sets: *lemma*, *choice*, *quantity*, *concept types*, *semantic dependencies*, *semantic head dependencies*, *question focus* and *question type*. By deriving a new feature subset (Section 4.2.1) and performing ablation test on features, a subset of features is derived which results in improved performance. The detail of all of these feature sets can be found in previous work [29]. After extracting eight feature sets, we found 4189 features to train the classification model. All of our classification algorithms use the same features for training the classification model. To further improve the classification results, we use the wrapper based feature selection method to select the best feature sets which could improve the classification performance [30]. Algorithm 1 shows the complete pseudo code of this process. The $F$ is all the input feature sets $\{f_1, f_1, \ldots, f_n\}$ and $S$ denotes the best selected sub-set.

**Algorithm 1.** Wrapper-based feature selection algorithm used for best feature subset selection

[4] https://github.com/wasimbhalli/Multi-label-Biomedical-QC-Corpus.

[5] https://github.com/clir/clearnlp.

[6] https://github.com/clir/clearnlp-bioinformatics-en-dep.

**Table 4**
Table illustrating number of questions in each question class after applying 65% threshold.

| Class | Questions | Class | Questions | Class | Questions | Class | Questions |
|---|---|---|---|---|---|---|---|
| umls:aapp | 213 | **choice** | 14 | umls:tmco | 4 | umls:humn | 1 |
| umls:gngm | 136 | umls:mbrt | 14 | umls:orgm | 4 | umls:dora | 1 |
| umls:dsyn | 116 | umls:celc | 13 | umls:inpr | 3 | tmtool:Species | 1 |
| umls:enzy | 69 | umls:biof | 13 | umls:comd | 3 | umls:grup | 1 |
| umls:bacs | 61 | umls:fndg | 12 | umls:emod | 3 | umls:phsf | 1 |
| umls:phsu | 59 | umls:bpoc | 12 | umls:amas | 3 | umls:antb | 1 |
| umls:sosy | 42 | umls:mobd | 10 | umls:fngs | 3 | umls:clna | 1 |
| umls:orch | 40 | umls:resa | 10 | umls:elii | 3 | umls:inpo | 1 |
| umls:qnco | 37 | umls:bact | 8 | umls:patf | 3 | umls:hlca | 1 |
| **quantity** | 34 | umls:lbpr | 8 | umls:tisu | 3 | umls:food | 1 |
| umls:clnd | 34 | umls:moft | 7 | umls:lbtr | 2 | umls:ftcn | 1 |
| umls:genf | 29 | umls:horm | 7 | umls:orga | 2 | umls:resd | 1 |
| umls:nusq | 24 | umls:phpr | 7 | umls:orgf | 2 | umls:popg | 1 |
| umls:rcpt | 20 | umls:qlco | 7 | umls:inch | 2 | umls:carb | 1 |
| umls:neop | 20 | umls:clas | 7 | umls:acty | 2 | umls:grpa | 1 |
| tmtool:Gene | 18 | umls:cnce | 7 | umls:blor | 2 | umls:medd | 1 |
| tmtool:Disease | 18 | umls:nnon | 7 | umls:geoa | 2 | umls:ortf | 1 |
| umls:mnob | 17 | umls:diap | 6 | umls:euka | 2 | umls:bdsu | 1 |
| umls:celf | 16 | umls:chem | 6 | umls:spco | 2 | tmtool:ProteinMutation | 1 |
| umls:topp | 16 | umls:cell | 5 | umls:npop | 2 | | |
| tmtool:Chemical | 15 | umls:virs | 5 | umls:bsoj | 1 | | |
| umls:imft | 14 | umls:cgab | 5 | umls:irda | 1 | | |

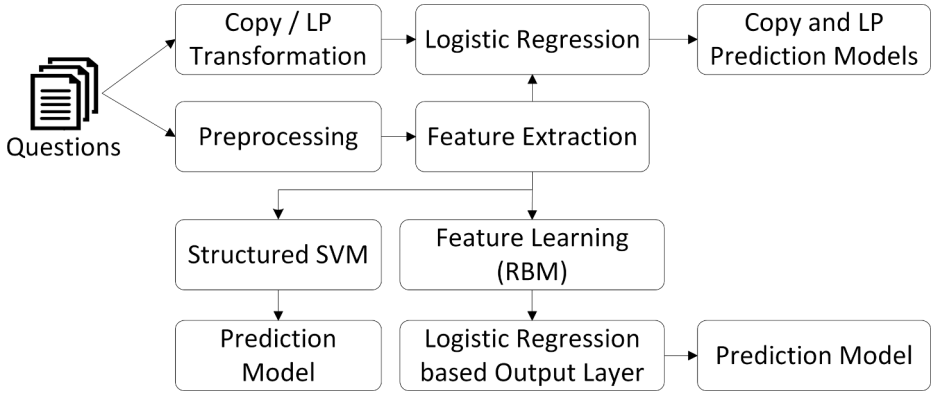


**Fig. 3.** Proposed multi-label biomedical question classification methodology.

**Input:** F
**Output:** S
$S \leftarrow \emptyset$
$score \leftarrow 0$
$optimalSubSetFound \leftarrow false$
**while** *!optimalSubSetFound* **do**
  **foreach** $f_i \in F$ **do**
    $score_{f_i} \leftarrow classification\_score_{(S \cup f_i)}$
  **end**
  $F_{max}, score_{max} \leftarrow arg\ max_i(score_{f_i})$
  **if** $score_{max} > score$ **then**
    $F \leftarrow F - \{F_{max}\}$
    $S \leftarrow \{F_{max}\}$
    $score \leftarrow score_{max}$
  **end**
  **else**
    $optimalSubSetFound \leftarrow true$
  **end**
  **return** S
**end**

We also performed experiments with focus driven semantic features. These semantic features are explained in next section.

#### 4.2.1. Focus-driven Semantic Features (FDSF)

We have already presented three types of semantic features, i.e. *concept type*, *semantic dependencies*, and *semantic head dependences* previously used in the OAQA system. We propose a set of focus-driven features from these semantic features. The focus driven subset will include only the semantic features derived from the question focus. For example, in question "which antibodies cause Riedel thyroiditis", only the *concept type*, *semantic dependencies*, and *semantic head dependences* related to the focus word *antibody* will be included in the feature space for classification. This subset not only reduces the dimensionality of feature space by 33% but also improves the classification performance as described in Results section (Section 5.4).

### 4.3. Data transformation

We use two data transformation techniques to transform multi-label questions to single-label so that logistic regression classification could be applied on it. These data transformation techniques are *copy* and *label power set* transformation. An example dataset (Table 5) is used in next paragraphs to explain both transformation techniques.

**Table 5**
Example dataset.

| Instances | Attributes | Label set |
|---|---|---|
| 1 | a1 | $\{\alpha_1, \alpha_3\}$ |
| 2 | a2 | $\{\alpha_4\}$ |
| 3 | a3 | $\{\alpha_1, \alpha_2, \alpha_3\}$ |
| 4 | a4 | $\{\alpha_1, \alpha_2\}$ |

**Table 6**
Data transformation using copy transformation method.

| Instances | Labels |
|---|---|
| 1a | $\alpha_1$ |
| 1b | $\alpha_3$ |
| 2 | $\alpha_4$ |
| 3a | $\alpha_1$ |
| 3b | $\alpha_2$ |
| 3c | $\alpha_3$ |
| 4a | $\alpha_1$ |
| 4b | $\alpha_3$ |

*Copy Transformation.* Copy transformation is a simple technique in which every multi-label instance is transformed into several instances, one per label. A variation of this method is dubbed copy transformation method which also assigns weights to the produced instances. These methods increase the total number of instances without losing any information [18]. Copy transformation method based data transformation from the example dataset of Table 5 to single label is shown in Table 6. Once transformed, the classification can be easily performed using the one-vs-all approach as used in a typical multi-class classification setting. As there are 85 classes in our dataset, we shall train one classifier for each class. Previously, the OAQA system used copy transformation technique for the task of multi-label question classification [23].

*Label Power Set Transformation (LP).* Label power set transforms the data in such a way that each distinct combination of labels available in the corpus acts as a unique class and then the single-label classification approaches are applied to it. Table 7 represents transformed dataset using label power set. Given the nature of the biomedical questions, the labels assigned to a single question are highly inter-dependent. For this reason, we have employed label power set for multi-label question classification. For our generated corpus, label power set transformation resulted in 210 unique merged labels, and we trained the same number of classifiers after applying this transformation.

### 4.4. Classification

We use three techniques for biomedical multi-label question classification:

- In transformation based methodology, we use logistic regression [31] after applying the data transformation techniques as described in Section 4.3.
- In adaptation based methodology, we use a modified version of SVM known as Structured SVM (SSVM) which does not require any data transformation technique and handles multi-label data directly. The details of SSVM can be found in the work of Thorsten Joachims [32].
- In deep learning based methodology, we use a Restricted Boltzmann Machine (RBM) with two layers (each containing ten neurons) for feature learning [33]. Moreover, we define an output layer having the same number of neurons as the number of labels in our dataset (85). The output layer contains logistic regression based classifier as activation function. In this model, the training is performed using contrastive divergence and backpropagation. Initially, weights are assigned using contrastive divergence which are later tuned using backpropagation. During forward propagation, the loss is calculated using the mean square error at the last layer. The sum of the loss across 85 output neurons is then back propagated using gradient descent to optimize the weights.

**Table 7**
Data transformation using label power set.

| Instances | Labels |
|---|---|
| 1 | $\{\alpha_{1,3}\}$ |
| 2 | $\{\alpha_4\}$ |
| 3 | $\{\alpha_{1,2,3}\}$ |
| 4 | $\{\alpha_{1,2}\}$ |

## 5. Experimental setup and results

This section describes the experimental setup used for multi-label question classification and the obtained results.

### 5.1. Experimental setup

In this section, we briefly describe online services and tools used for the experimentation of question classification. First of all, question corpus is generated by the process as described in Section 3. To generate question corpus, we used OAQA's internal system.[7] This system utilizes services of two online tools known as tmtool and UTS. TmTool provides a RESTful API that is used to annotate biomedical concepts.[8] It incorporates four tools which are tmVar,[9] tmChem,[10] DNorm,[11] and GNormPlus.[12] tmVar is a text-mining mutation information annotation tool whereas tmChem identifies all types of chemicals. DNorm is used to identify diseases, and GNormPlus is a gene/protein concept recognizer. Using these four tools incorporated into tmtool, we can annotate most of the possible biomedical concepts present as an answer in BioASQ dataset.

On the other hand, UTS is a UMLS Terminology Service that is used to search and display contents from UMLS.[13] UMLS is a comprehensive biomedical database and ontology that is available in various languages. It returns semantic types and semantic relationships between different biomedical terms using three knowledge resources, i.e. MetaThesaurus, Semantic Network, and SPECIALIST Lexicon. In our case, we are using UTS service to identify biomedical concepts and annotating them using their relevant semantic types.

The features are extracted using OAQA setup, and we save them in Weka compatible format.[14] To evaluate the copy and label set transformation, we use Meka[15] which is an alternative to Weka developed for multi-label classification. Similarly, we use the Meka tool to perform classification using Restricted Boltzmann Machine (RBM). We linearly stack two RBMs to define our model to extract discriminative features - the hidden layer of the first RBM act as a visible layer of second RBM. For pre-training of RBM, we use contractive divergence. To further enhance the performance of the model, we back-propagate the error to adjust weights of RBM layers. The visible layer of the first RBM comprises of 4189 units which are equal to the number of input features. The hidden layer of both RBMs consists of ten hidden units. We set the value of learning rate and momentum to 0.1 during experimentation,

[7] https://github.com/oaqa/bioasq.
[8] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/.
[9] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/.
[10] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmchem/.
[11] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm/.
[12] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/.
[13] https://uts.nlm.nih.gov/home.html#.
[14] https://www.cs.waikato.ac.nz/ml/weka/arff.html.
[15] http://meka.sourceforge.net.

and train the model for 1000 epochs. For logistic regression based classification, LibLinear tool was utilized [31]. To perform experiments for structured support vector machine (SSVM), we use structpy python library[16] with *unary* structured model having *regularization* and *convergence tolerance* values as 0.1 and 0.01 respectively. All the methods are evaluated using 10-fold cross-validation.

### 5.2. Datasets

As already discussed in Section 3.2, we developed our dataset to perform experiments on question classification. To evaluate the performance of the complete QA system, we use the final batch of BioASQ dataset for the 5*th* year which contains 35 factoid questions.[17]

### 5.3. Performance metrics for multi-label classification

Multi-label classification evaluation measures are broadly divided into two categories namely "Label Based" and "Example-Based" [34]. Label based evaluation measures are considered the modified forms of single-label evaluation measures as they first calculate the performance of each label separately and then average the performance of all labels. Whereas, example-based evaluation measures are specially built for multi-label classification because they deal with the average values of predicted and actual labels over all the examples of a corpus. We have exploited example based measures in our experimentation namely *Micro* $F_1$, *Accuracy*, and *Hamming Loss*. The details of these techniques can be found in previous studies [20,35].

### 5.4. Results

In this section, we present the results of biomedical question classification. We also evaluate the performance of the proposed question classification process when integrated with a complete QA system.

#### 5.4.1. Results of question classification for LAT prediction

This section reports the results of LPLR with other state-of-the-art techniques namely Restricted Boltzmann Machine (RBM), Structured Support Vector Machine (SSVM), and CLR (previously used by OAQA system). We utilize a set of evaluation measures(Micro $F_1$ Measure, Accuracy, Hamming Loss) to assess the performance of these techniques over a set of 8 features (*Lemma, Choice, Quantity, Question Type, Question Focus, Concept Type, Semantic Dependencies, Semantic Head Dependencies*). Figs. 4–6 show the performance of all aforementioned techniques in terms of $F_1$ Measure, Accuracy and Hamming Loss respectively.

As Fig. 4 suggests, copy transformation based logistic regression model and SSVM are on approximately the same performance level with 1% difference. Whereas, the LPLR exceeds the CLR with a margin of 4%. On the other hand, RBM shows a marked difference from the other three methods with only a score of 0.28. So concerning $F_1$ Measure, LPLR has surpassed all the other multi-label classification techniques.

Likewise, similar sort of performance trend is observed in another performance evaluation metric named as Accuracy. As shown in Fig. 5, CLR still shows around 2.5% better performance as compared to SSVM. As with $F_1$ Measure, RBM performance again decreases with a visible difference of 21% from SSVM. Contrary, LPLR again transcends other techniques with a figure of 0.42.

As Fig. 6 illustrates, the performance of copy transformation based logistic regression is slightly better than other classification techniques showing hamming loss of 0.018. LPLR and SSVM shows same figure of 0.02 and RBM shows unexpectedly minor loss of 0.019 with regard of its performance with other measures.

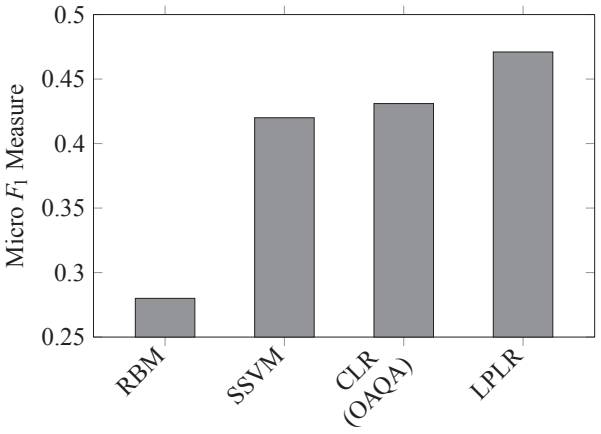


**Fig. 4.** Performance comparison of label power set based logistic regression (LPLR) with RBM, Structured SVM (SSVM), and copy based logistic regression (CLR) with respect to Micro $F_1$ Measure.

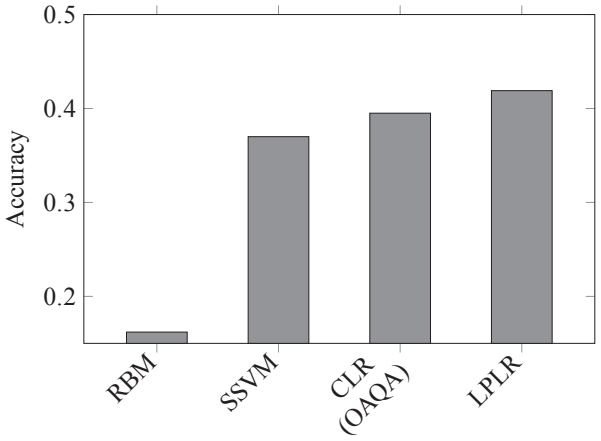


**Fig. 5.** Classification accuracy comparison of label power set based logistic regression (LPLR) with RBM, Structured Support Vector Machine (SSVM) and copy based logistic regression (CLR).

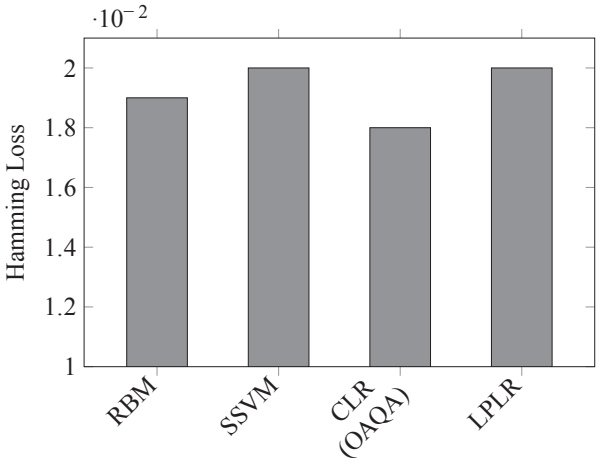


**Fig. 6.** Hamming loss comparison of label power set based logistic regression (LPLR) with CLR, RBM, and SSVM.

To evaluate the effect of selected feature sets on multi-label classification, we performed an ablation test as shown in Table 8. We use all the eight feature sets described in Section 4.2. On the other hand, feature-driven semantic features (FDSF) are the modified feature sets as described in Section 4.2.1. We assessed each feature individually and

[16] https://pystruct.github.io/.
[17] http://participants-area.bioasq.org/Tasks/b/testData/.

**Table 8**
*$F_1$Measure* and *Accuracy* results on incremental combination of baseline and focus-driven semantic features (FDSF); L:Lexical, F:Focus, Q: Quantity, C: Choice, CT: Concept Type, QT: Question Type, SD: Semantic Dependencies, SHD: Semantic Head Dependencies, AP: All Previous.

| Feature subset | | *$F_1$Measure* | | Accuracy | | *$\Delta F_1$measure* | | Δ accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | FDSF | Baseline | FDSF | Baseline | FDSF | Baseline | FDSF | Baseline | FDSF |
| L | L | 0.45 | 0.45 | 0.41 | 0.41 | | | | |
| L + F | L + CT | 0.48 | 0.48 | 0.43 | 0.43 | 0.03 | 0.03 | 0.02 | 0.02 |
| AP + Q | AP + Q | 0.49 | 0.49 | 0.44 | 0.44 | 0.01 | 0.01 | 0.01 | 0.01 |
| AP + C | AP + C | 0.49 | 0.49 | 0.45 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 |
| AP + CT | AP + QT | 0.49 | 0.49 | 0.44 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 |
| AP + QT | AP + SD | 0.49 | 0.49 | 0.44 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 |
| AP + SD | **AP + F** | 0.48 | **0.50** | 0.43 | **0.45** | −0.01 | 0.01 | −0.01 | 0.01 |
| AP + SHD | AP + SHD | 0.47 | 0.50 | 0.43 | 0.44 | −0.01 | 0.00 | −0.01 | 0.00 |

**Table 9**
Performance comparison of copy with our introduced label power set (LP) transformation technique on the final batch of 5*th* year BioASQ challenge - (Δ represents the performance gain).

| Strict accuracy | | | Lenient accuracy | | | MRR | | |
|---|---|---|---|---|---|---|---|---|
| CLR | LPLR | Δ | CLR | LPLR | Δ | CLR | LPLR | Δ |
| 0.11 | 0.14 | 0.03 | 0.20 | 0.26 | 0.06 | 0.14 | 0.19 | 0.05 |

found that lexical feature played a vital role in the performance of question classification. We used two different iterations for combining the previously defined eight features: (1) we used the feature sets already used in OAQA which we termed as the baseline, and (2) on similar features with focus-driven semantic features (FSDF). Similar to wrapper method discussed in Section 4.2, we incrementally added the best performing feature sets to the previous feature set and found that, in the case of baseline, the lexical and focus feature showed the best performance. In the case of FSDF, the combination of lexical with focus performed similarly as the combination of lexical features and concept types. The table shows that focus driven semantic features achieved the best performance without considering the semantic head dependencies feature set. The best performing features are shown in bold.

#### 5.4.2. Effect of LAT prediction performance on biomedical QA system

The paper in hand focuses on the performance enhancement of biomedical LAT prediction. However, to evaluate the effect of better LAT prediction on QA system, in this section, we present the results of factoid answers which are assessed using strict accuracy, lenient accuracy, and mean reciprocal rank (MRR) [36].

The results in Table 9 show that, for all evaluation measures, the performance of the LPLR surpassed the performance of CLR which improved the performance of factoid question answering.

## 6. Discussion

This section summarizes the issues faced during the dataset generation and present the factors which are responsible for the better performance of label power set as compared to the previously used copy transformation technique. We also draw highlights of findings in feature subset selection and the ablation test performed for the selection of high performing features.

The corpus generation process in OAQA assigns no labels to questions for which any semantic type is not found. For a particular question, if the unknown entities are higher than the specified threshold, *unlabeled* type is added. For 772 out of 899 questions were either only assigned *unlabeled* as their lexical answer type or *unlabeled* was included with other assigned labels. Due to this reason, the *unlabeled* type was always predicted, and it showed a strong association with all other labels thus affecting the performance of all the classification methodologies. So, we excluded all questions having only the null type as their LAT, and for questions where null was included with different types for a question, the *unlabeled* type was removed during the corpus generation process.

LPLR overshadowed the performance of CLR with a huge margin as label power set assists the classifier to model label cardinality and correlations effectively during training. Whereas, copy transformation generates an extensive dataset with high label cardinality and it provides no obvious way to model label dependencies in multi-label data transformation. The reason behind the lousy performance of copy transformation based logistic regression is that copy transformation creates the same instances of data with different classes and ultimately produce confusions for the classifier in defining the decision boundary. So, two distinct labels assigned to the same question will have the same input features, but the output label will be different. So repeated features with varying class labels lose the strength of their discrimination among different classes.

To analyze the drawbacks of copy transformation approach more effectively, let us consider the feature (Custom defined feature types). Custom defined feature types are assigned to only those questions which have no semantic type assigned by UTS or TmTool, and they contain either of the following features (start with do, is/be or contain words like how many, how much, or, rate, number). Custom defined feature types assign the value of 0 or 1 depending on the absence or presence of different words. For instance, consider the two following questions which are assigned custom defined feature types (choice, quantity) based on features discussed above.

1. choice: **Do** archaeal genomes contain one **or** multiple origins of replication?
2. quantity: What is the **rate** of survival after commotio cordis?

Custom defined feature types have utilized following features (Do, or, rate) to assign specific semantic type (choice or quantity) against each mentioned question. As custom defined feature types heavily depend on particular features (e.g. do, or, rate), specific biomedical questions satisfy such rules, but they do not belong to custom defined feature types. For example, consider the following questions:

1. umls:dsyn,umls:gngm Which is the most common disease attributed to malfunction **or** absence of primary cilia?
2. umls:euka,umls:fngs,umls:dsyn List **invertebrates** where ultra-conserved elements have been identified.

This phenomena badly affects the decision boundary of a classifier in case of copy transformation where multiple copies of the same question are created, one per label assigned to it. However, due to the nature of the label power set approach, it does not produce such issues for the underlying classifier. Label power set based logistic regression also predicts a set of labels with random length, so there is no need to define a fixed length for predicted labels.

On the other hand, RBM and SSVM showed poor performance compared to logistic regression with transformation techniques such as copy and label power set. We concluded that the performance of RBM is highly dependent on the characteristics of data. Deep learning based feature extraction in RBM improves performance by extracting and learning better dependence relationships between discriminative features and labels. In the case of multi-label document classification when features are modeled using RBM, it offers better performance on datasets such as Music, Scene and Yeast [22]. However, multi-label question classification is an entirely different process as here we have only a small set of features and label cardinality. Similarly, SSVM does not incorporate label cardinality as it deals with the number of labels as a whole. The SSVM algorithm decides the boundary based on the question content (discriminative features) and their relevance with class labels. Similar to RBM, in a multi-label paradigm it works better on large textual documents because of large feature space.

In the case of feature engineering, we observed that in feature-driven semantic features (FDSF), a semantic feature's contribution was below when compared with similar features in the baseline. However, combining it with other features improved the classification results. Similarly, for focus-driven feature subset the quantity, choice, and question type did not contribute to the classification performance. Adding semantic dependencies did not have any effect on the FDSF. However, it decreased the $F_1$ score when combined with a subset of baseline features. Semantic head dependencies affected both accuracy and $F_1$ measure in baseline and hurt accuracy measure for FDSF where it decreased the performance by 1%. It shows that semantic dependency is not a very useful feature and does not provide any additional performance advantage even when used in FDSF. Quantity and choice contribute individually low results, but when combined with other features, it improves overall classification results. On the other hand, although semantic dependencies and semantic head dependencies individually provide better classification results as compared to choice and quantity, their performance degrades when combined with other features.

A unique subset of seven new features (excluding semantic head dependencies) with label power set resulted in classification performance of 0.50, and o.44 regarding $F_1$*Measure* and *Accuracy* respectively. It shows that focus driven semantic features when combined with other features not only facilitate in dimensionality reduction but also improves biomedical question classification.

## 7. Conclusion and future work

In this paper, we developed a biomedical LAT corpus (MLBioMedLAT) by combining the strengths of the OAQA system [29] and BioMedLAT [12]. We used the first four years BioASQ dataset prepared by biomedical experts for the task of biomedical question answering. We introduced LPLR for biomedical question classification which improved the performance as compared to the previously used CLR with a margin of 4%. We also proposed a focus-driven feature subset which, combined with LPLR, improved the classification performance by 7%. We also compared LPLR with two other introduced techniques in biomedical question classification domain namely, RBM and SSVM with the technique previously used (CLR). We evaluated the integrity of these techniques through a set of performance metrics (F1 Measure, Accuracy, Hamming Loss) over baseline and a set of eight features (Lemma, Choice, Quantity, Concept Types, Semantic Dependencies, Semantic Head Dependencies, Question Type, Question Focus). We found that label power set outperformed all other techniques both regarding $F_1$ and accuracy measure whereas for the hamming loss it showed results almost similar to other techniques. Lexical answer type prediction for multi-labeled questions poses challenges for classifiers as it is challenging to find a decision boundary. Furthermore, such a dataset is highly skewed. Such problems must be solved to advance the performance of biomedical question classification performance leading to improved question answering systems. The use of null type or unknown type is also an exciting aspect which needs further exploration, and there is a necessity of better evaluation measures if an unknown type is to be included in the dataset. There is also a desperate need to introduce and improve more data transformation and algorithm adaptation techniques to enhance the performance of question classification in the biomedical domain. We also need more advanced feature engineering and feature generation techniques to extract highly discriminative features for multi-label learning. In the future, we shall work on these aspects to further enhance biomedical LAT prediction performance.

## Conflict of interest

None declared.

## Appendix A. Corpus generation

Let $C_u$ be the unlabeled corpus consists of different questions. $q_i$ represents the *ith* question in the corpus where the range of $i$ is between 1 and 899. To get semantic types of questions from UTS we have to pass question and *Mode* as parameters to the UTS function where *Mode* is defined as *Mode* = {'*exact*', '*word*'}.

**Algorithm 2.** Conversion of unlabeled corpus to labeled corpus

**Input:** C
**Output:** LabeledCorpus
initialization of **UnlabeledToLabeled** function;
**foreach** *question* $q_i$ *in* $C_u$ **do**
  $exactAnswertype_i \leftarrow$ searchInGold($q_i$) ;
  $utsSemanticTypes_i \leftarrow$ UTS($exactAnswertype_i$, Mode);
  $tmtoolSemanticTypes_i \leftarrow$ tmtool($exactAnswertype_i$) ;
  $combined_i \leftarrow$ Combine($utsSemanticTypes_i$, $tmtoolSemanticTypes_i$ );
  $bioMedLAT_i \leftarrow$ getBioMedLAT($q_i$)
  **if** $combined_i$ *contains* $bioMedLAT_i$ **then**
    $LabeledCorpus_i \leftarrow$ threshold($combined_i$,
     getConfidence($bioMedLAT_i$));
    continue;
  **end**
  **if** $combined_i$ != *empty* **then**
    // Add bioMedLAT with confidence score of 1;
    $combined_i \leftarrow$ Combine($bioMedLAT_i$, $combined_i$ );
    $LabeledCorpus_i \leftarrow$ threshold($combined_i$, 0.65);
  **else**
    $LabeledCorpus_i \leftarrow$ customDefineType($q_i$) ;
    **if** $LabeledCorpus_i$ = *empty* **then**
      $LabeledCorpus_i = bioMedLAT_i$;
    **end**
  **end**
**end**
return removeUnlabeledClass(LabeledCorpus);

initialization of **customDefineType** function;
**Input:** Question
**Output:** Custom Defined Semantic Type
**if** ($q_i$ *starts with "do")* | | ($q_i$ *starts with "is")* | | ($q_i$ *starts with "be")*) | | ($q_i$ *contains "or")* **then**
  **return** *"choice"*;
**end**
**if** ($q_i$ *contains "how many")* | | ($q_i$ *contains "how much")* | | ($q_i$ *contains "how large")*| | ($q_i$ *contains "how long")*| | ($q_i$ *contains "diameter")*| | ($q_i$ *contains "value")*| | ($q_i$ *contains "rate")*| | ($q_i$ *contains "percentage")*| | ($q_i$ *contains "incidence")* | | ($q_i$ *contains "prevalence")*| | ($q_i$ *contains "proportion")*| | ($q_i$ *contains "number")* **then**
  **return** *"quantity"*;
**else**
  **return** *"unlabeled"*;
**end**

## Appendix B. Data annotation services

The services used for labeling biomedical questions are briefly described in the following sections.

### *B.1. UMLS Terminology Service (UTS)*

The Unified Medical Language System (UMLS) terminology service (UTS) is a gateway which provides web-based access to UMLS knowledge sources [37]. The service is based on a client–server architecture where clients can send requests to the knowledge source servers managed by the U.S. National Library of Medicine.[18] The UMLS metathesaurus used in UTS contains the largest biomedical thesaurus which is organized by concepts and provides a link to similar names in nearly two hundred different vocabularies. The UMLS semantic network offers 133 categories or semantic types for a consistent categorization of all concepts present in UMLS metathesaurus. These semantic types have been further categorized into 14 semantic groups or coarse classes to reduce the complexity of 133 categories [38]. UTS may be queried to retrieve semantic types for biomedical entities.

[18] https://www.nlm.nih.gov/.

### B.2. Tmtool

Tmtool is web-based text mining service for identifying biomedical concepts [39]. The service integrates several state-of-the-art tagging systems (DNorm, GNormPlus, SR4GN, tmChem, and tmVar) and offers arbitrary text input to identify biomedical concepts as shown in Fig. 7. The details of these individual tagging systems may be found in [40–44].

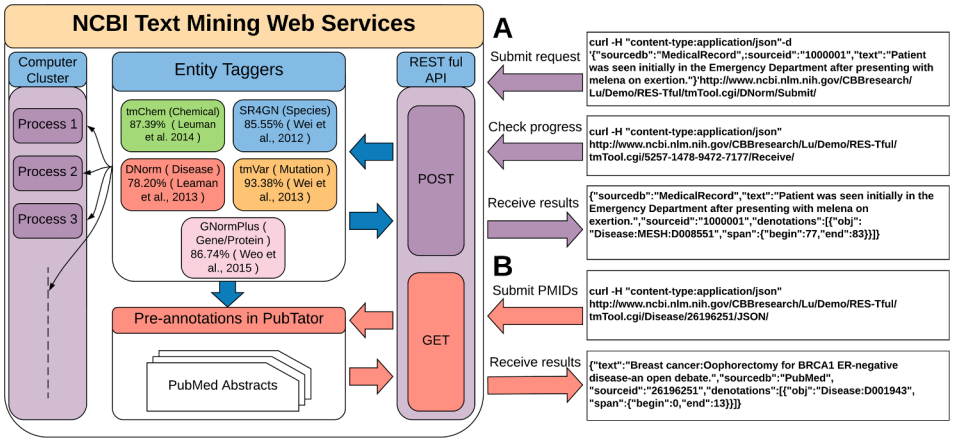


**Fig. 7.** TmTool Service Architecture [39].

## Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103143.

## References


[1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC Bioinform. 16 (1) (2015) 1.

[2] D. Jurafsky, J.H. Martin, Speech and Language Processing (Prentice Hall Series in Artificial Intelligence).

[3] R.F. Simmons, Answering english questions by computer: a survey, Commun. ACM 8 (1) (1965) 53–70.

[4] W.A. Woods, Progress in natural language understanding: an application to lunar geology, Proceedings of the June 4–8, 1973, National Computer Conference and Exposition, ACM, 1973, pp. 441–450.

[5] W.G. Lehnert, A conceptual theory of question answering, Proceedings of the 5th International Joint Conference on Artificial Intelligence-Volume 1, Morgan Kaufmann Publishers Inc., 1977, pp. 158–164.

[6] E.M. Voorhees, D.K. Harman, The Eighth Text Retrieval Conference (trec-8), Tech. Rep., 2000.

[7] D. Mollá, J.L. Vicedo, Question answering in restricted domains: an overview, Comput. Linguist. 33 (1) (2007) 41–61.

[8] W. Hersh, E. Voorhees, Trec Genomics Special Issue Overview, 2009.

[9] M. Sarrouti, A. Lachkar, S.E.A. Ouatik, Biomedical question types classification using syntactic and rule based approach, Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on, vol. 1, IEEE, 2015, pp. 265–272.

[10] D. Weissenborn, G. Tsatsaronis, M. Schroeder, Answering Factoid Questions in the Biomedical Domain, BioASQ@ CLEF 1094.

[11] Y. Zhang, S. Peng, R. You, Z. Xie, B. Wang, S. Zhu, The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering, in: CEUR Workshop Proceedings, CEUR Workshop Proceedings, vol. 1391, 2015.

[12] M. Neves, M. Kraus, Biomedlat Corpus: Annotation of the Lexical Answer Type for Biomedical Questions, OKBQA 2016, 2016, p. 49.

[13] M. Zhou, F. Wei, X. Liu, H. Sun, Y. Duan, C. Sun, H.-Y. Shum, Learning-based Processing of Natural Language Questions, uS Patent App. 13/539,674, January 2 2014.

[14] J. Silva, L. Coheur, A.C. Mendes, A. Wichert, From symbolic to sub-symbolic information in question classification, Artif. Intell. Rev. 35 (2) (2011) 137–154.

[15] X. Li, D. Roth, Learning question classifiers, Proceedings of the 19th International Conference on Computational Linguistics, vol. 1, Association for Computational Linguistics, 2002, pp. 1–7.

[16] Z. Huang, M. Thint, Z. Qin, Question classification using head words and their hypernyms, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 927–936.

[17] V. Roth, B. Fischer, Improved functional prediction of proteins by learning kernel combinations in multilabel settings, BMC Bioinform. 8 (2) (2007) S12.

[18] E. Gibaja, S. Ventura, A tutorial on multilabel learning, ACM Comput. Surv. (CSUR) 47 (3) (2015) 52.

[19] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, Data mining and Knowledge Discovery Handbook, Springer, 2009, pp. 667–685.

[20] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.

[21] A. Clare, R.D. King, Knowledge discovery in multi-label phenotype data, European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2001, pp. 42–53.

[22] J. Read, F. Perez-Cruz, Deep learning for multi-label classification. Available from: arXiv preprint arXiv:1502.05988.

[23] Z. Yang, Y. Zhou, E. Nyberg, Learning to answer biomedical questions: Oaqa at bioasq 4b, in: Proceedings of the Fourth BioASQ workshop, 2016, pp. 23–37.

[24] D. Metzler, W.B. Croft, Analysis of statistical question classification for fact-based questions, Inform. Retr. 8 (3) (2005) 481–504.

[25] R. Bunescu, Y. Huang, Towards a general model of answer typing: Question focus identification, in: Proceedings of The 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010), RCS Volume, 2010, pp. 231–242.

[26] F. Schulze, R. Schüler, T. Draeger, D. Dummer, A. Ernst, P. Flemming, C. Perscheid, M. Neves, Hpi question answering system in bioasq 2016, in: Proceedings of the Fourth BioASQ workshop, 2016, pp. 38–44.

[27] X. Li, D. Roth, Learning question classifiers: the role of semantic information, Nat. Lang. Eng. 12 (3) (2006) 229–249.

[28] A. Blum, Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain, Mach. Learn. 26 (1) (1997) 5–23.

[29] Z. Yang, N. Gupta, X. Sun, D. Xu, C. Zhang, E. Nyberg, Learning to answer biomedical factoid & list questions: Oaqa at bioasq 3b, in: CLEF (Working Notes), 2015.

[30] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.

[31] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (Aug) (2008) 1871–1874.

[32] T. Joachims, Support Vector Machine for Complex Outputs, 2018 (Online; accessed 1-Sept-2018). < https://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html > .

[33] G. Hinton, S. Osindero, M. Welling, Y.-W. Teh, Unsupervised discovery of nonlinear structure using contrastive backpropagation, Cogn. Sci. 30 (4) (2006) 725–731.

[34] M.S. Sorower, A Literature Survey on Algorithms for Multi-label Learning, Oregon State University, Corvallis 18.

[35] M.N. Asim, A. Rehman, U. Shoaib, Accuracy based feature ranking metric for multi-label text classification, Int. J. Adv. Comput. Sci. Appl. 8 (10) (2017) 369–378.

[36] G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, E.

Gaussier, P. Gallinari, T. Artieres, M.R. Alvers, M. Zschunke, et al., Bioasq: a challenge on large-scale biomedical semantic indexing and question answering, in: AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text, 2012.
[37] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucl. Acids Res. 32 (suppl_1) (2004) D267–D270.
[38] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating umls semantic types for reducing conceptual complexity, Stud. Health Technol. Inform. 84 (0 1) (2001) 216.
[39] C.-H. Wei, R. Leaman, Z. Lu, Beyond accuracy: creating interoperable and scalable text-mining web services, Bioinformatics 32 (12) (2016) 1907–1910.
[40] C.-H. Wei, H.-Y. Kao, Z. Lu, Gnormplus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains, BioMed Research International, 2015.
[41] R. Leaman, C.-H. Wei, Z. Lu, tmchem: a high performance approach for chemical named entity recognition and normalization, J. Cheminform. 7 (1) (2015) S3.
[42] R. Leaman, R. Islamaj Doğan, Z. Lu, Dnorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (22) (2013) 2909–2917.
[43] C.-H. Wei, H.-Y. Kao, Z. Lu, Pubtator: a web-based text mining tool for assisting biocuration, Nucl. Acids Res. 41 (W1) (2013) W518–W522.
[44] C.-H. Wei, H.-Y. Kao, Z. Lu, Sr4gn: a species recognition software tool for gene normalization, PloS One 7 (6) (2012) e38460.