ISSN 2319-2720

Volume 9, No.2, April - June 2020

International Journal of Computing, Communications and Networking

Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn02922020.pdf

https://doi.org/10.30534/ijccn/2020/02922019

# Prediction of Question Tags Based on LDA and Deep Neural Network

**Ashwin Cherry Mathew**[1], **Sreenimol K. R**[2]

[1]PG Student, Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kerala, India, amathew790@gmail.com

[2]Associate Professor, Department of Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kerala, India, sreenimol.kr@mangalam.in

## ABSTRACT

Students are being evaluated based on the examinations conducted by various institutions or organizations, which test the knowledge of that person. Exams like Computerized Adaptive Testing (CAT), offers a computer based test that adapts the examinee's ability level. Some of the CAT exams include tags which help students to understand the questions. Tags are metadata used to identify or describe an item. There are three types of tags: Manual Tagging, Semi-Automatic tagging and Fully Automatic tagging. Earlier manual tagging was used to construct question banks. However it is time consuming and leads to many other consistency issues. A Semi-Automatic tagging facilitates human intervention to increase the accuracy of tagging. Fully automatic tagging gives a more promising result as compared with manual and semi-automatic tagging. This paper proposes a fully automated tagging system which uses Deep Neural Network and Natural Language Processing to generate tags from the derived knowledge unit. This paper also discusses LDA (Latent Dirichlet Allocation) which gives the relevance of each tag.

**Key words: Natural** Language Processing, Latent Dirichlet Allocation (LDA), Deep Neural Network, Text Analysis.

## 1. INTRODUCTION

Internet has influenced a lot in areas like social-medias, banking sectors, education and many other communication platforms. The rapid growth of internet had turned many handwritten exams to computerized tests. These test helped the examiner to measure the ability and to improve learning skills of each student. Computerized Adaptive Testing (CAT) allows to progressively adjusting the approach and provides a recommendation which will improve their learning efficiently. A well-organized and structured question bank is stored in the database (usually multiple choice questions). Also, a knowledge map is provided for effective knowledge management tool which can produce knowledge unit. Tags are provided with the question in order to help or understand the question and provide efficient way to organize resource. The knowledge units can be used as tags which utilize tagging to associate questions. Knowledge units are extracted after analyzing the knowledge map.

Tagging can be done by three types: (1) Manual Tagging, (2) Semi-automatic Tagging and (3) Automatic Tagging.

- Manual Tagging is one of the main and earliest methods that organize questions. These are mostly performed using handwritten tags. However there are few drawbacks for this method. First, the taggers need good knowledge of the subject in order to tag the questions using knowledge map. Second, manual tagging requires more time and cost for managing and updating the questions within a question bank. Different taggers analyze questions in a different perspective, which will lead to a different result.

- Semi-Automatic Tagging analyzes questions and return the tags which then processed by users to make the final approval. Human help is mandatory. Although it gives a more accurate result than manual tagging, it takes more time to finalize.

- Automatic Tagging does not need human intervention, it gives a standard and consistent result with less amount of time and cost. The tags are produced based on knowledge unit by analyzing the knowledge map.

This paper presents an automatic question tagging which uses the concept of Deep Neural Network to generate the appropriate tag. A question bank containing a variety of questions, in which it processes and generate tags that corresponds to a particular subject or module within a particular subject. A knowledge Map is created to identify each subject based on the knowledge unit. Text Mining is used to analyze the questions and to study the pattern of the questioning. NLP is also used to mine each questions based on their English grammar.

## 2. EXISTING SYSTEM

There are many techniques in which automatic tagging can be achieved using machine learning concepts like classification models or topic models.

In [2] and [7], discuss the tagging in Stack Overflow site. Stack Overflow is a question and answer site which helps to clarify doubts about various programming topics. This paper suggests the functioning of stack overflow site. It uses a tf-idf (Term frequency and Inverse Document Frequency) method to describe the knowledge unit. It also uses a Naïve Bayes classifier which classifies and predicts tags based on the tf-idf. A TagStack system is used to tag Q&A services.

A detailed study of tf-idf is explained in [3]. Tf-Idf weighting, stands for *term frequency (tf) inverse document frequency (idf*). Tf-idf weighting is used mainly in text mining and information retrieval. They are used to reveal the linguistic terms in a studied corpus. Term importance increases with the term's frequency in the text, yet is offset by the frequency of the term in the domain of interest.

Given a collection of terms $t \in T$, which appear in N document is $d \in D$ of length $n_d$, tf-idf weighting is computed as follows

$$Tf_{t,d} =$$

$$Idf_t = \log$$

$$Wt,d = tf_{t,d} \quad idf_t$$

where$f_{t,d}$ is the frequency of term $t$ in document $d$, and $df_t$ is the document frequency of term $t$, that is, the number of documents in which term $t$ appears. Text mining can be also used for applications like detecting criminal activities in social media platforms as mentioned in [9].

Natural Language Processing holds a wide range of areas such as Text translation, Human-computer dialogue, Information retrieval, Natural Language Understanding and so on. These areas can converted into four parts: Linguistic orientation, data processing orientation, artificial intelligence and scientific cognitive orientation and language engineering orientation. A question answering system is explained in [4], which uses Natural Language Processing. The system consist of three components: (1) Problem processing where the user submits the question which will be analyzed and filtered in natural language form, (2) Information retrieval where it analyses the question type is and required answer is mentioned. (3) Answer extraction where the answer will be extracted based on the keywords and their extensions. The extraction of keywords in question sentences has a great probability that will affect the effect of information retrieval. The search keywords are mainly nouns, verbs, adjectives and so on, and the keywords do not contain the interrogative words in question sentences. The optimizing method of this question answering system can be shown in Figure 1. A Web monitoring system is an example for NLP. This system [8] controls the sensitive information spreading on intent

Latent Dirichlet Allocation in [5], is an example for topic modelling. Latent Dirichlet allocation (LDA) is a probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. [10] shows how LDA and Deep Learning can be used for feature extraction and sentiment analysis. LDA extracts topics and uses deep learning for extracting topics for integration. Nowadays Business related data (which include Feature Extraction in Text and Data Mining for Financial Time Series Prediction) extraction can be done using LDA and such kind of LDA is called FinLDA[11].

Latent Semantic Analysis in [6], explains as it is a method for analyzing text with certain mathematical computation and analyzing relationship between terms in the document. The importance of LSA in a document is to find hidden relationship in document and understand the relationship between terms in that document.

With the above existing systems can be helpful in creating the automatic question tagging. As this paper deals with tagging the questions with tf-idf, Natural Language Processing using Deep Neural Network.

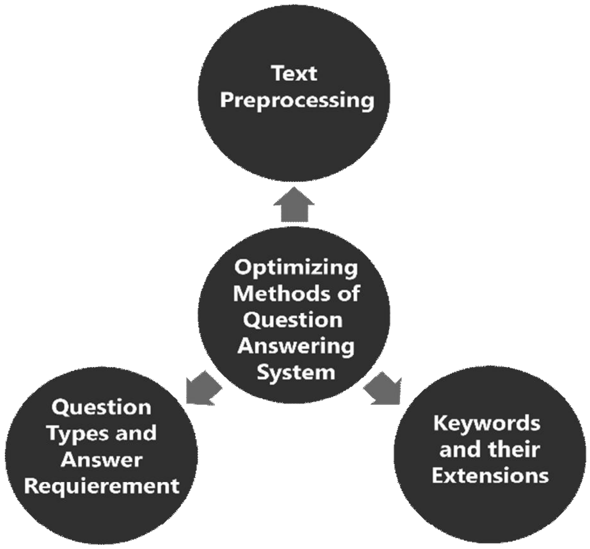

Text Preprocessing
Optimizing Methods of Question Answering System
Question Types and Answer Requierement
Keywords and their Extensions

**Figure 1**: Optimizing Methods

## 3. PROPOSED SYSTEM

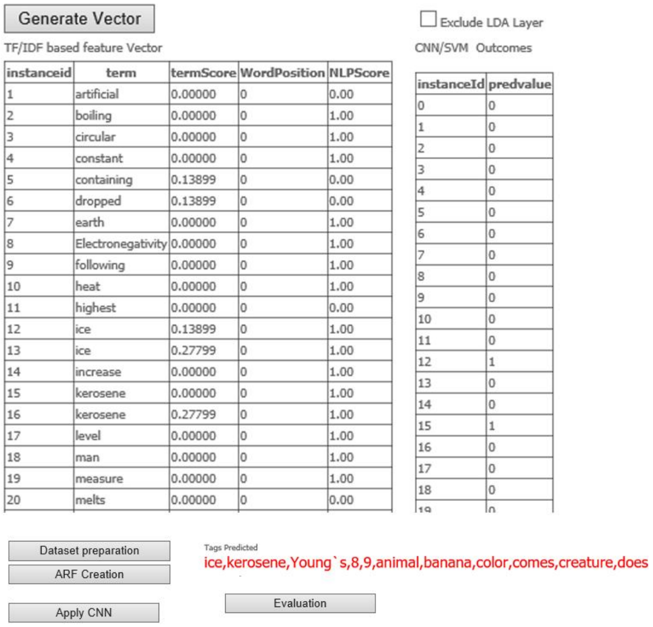
Automatic question tagging using deep neural network helps to tag the questions based on knowledge map[1]. The main purpose of this tagging system is text mining where the questioner can understand which all questions belongs for each subject. For that, a question bank is created which can include questions of different subjects. These questions can be stored in a database for further review.

The words within a question are to be break down into tokens. Each token may be a stopword or a knowledge unit. Actually stopwords are words which are filtered out before or after processing of Natural Language Processing. So the stopword removal is essential for creating knowledge map. Knowledge units are those which are later used for text mining. The subjects are selected based on these knowledge maps.

After removing stopwords, the next step is to assign each of the remaining words with tf-idf value. The numerical value indicates how important a word is to a document (here in a question bank). In general it is used as the weighing factor in searches of information retrieval, text mining, and user modeling. Tf-idf value increases the probability as the number of times a word appears in a document.

Convolution Neural Network allows each neuron to accept an input, performs an operation (dot product) and optionally follows it with a non-linearity to produce an output. ConvNet is basically used for detecting patterns. Each of these patterns are detected by filters. This allows to easily identify the knowledge unit as it is passed on to the deeper layers.

Latent Dirichlet Allocation is a probabilistic model for collections of discrete data such as text corpus (which contains resources of large set of texts). LDA is a topic model that generates topic based on word frequency from a set of documents. Here, the tags are generated based on this topic model. LDA finds accurate mixture of topics within a given question. The question bank may contain questions from different subjects where LDA separates each question based on the subject. So, each topic is modeled as infinite mixture over a set of probabilities. For text modelling, the topic probabilities provide an explicit representation of document. This concept is used to extract better and more powerful tags from the Question corpus.



TF/IDF based feature Vector

| instanceid | term | termScore | WordPosition | NLPScore |
|---|---|---|---|---|
| 1 | artificial | 0.00000 | 0 | 0.00 |
| 2 | boiling | 0.00000 | 0 | 1.00 |
| 3 | circular | 0.00000 | 0 | 1.00 |
| 4 | constant | 0.00000 | 0 | 1.00 |
| 5 | containing | 0.13899 | 0 | 0.00 |
| 6 | dropped | 0.13899 | 0 | 0.00 |
| 7 | earth | 0.00000 | 0 | 1.00 |
| 8 | Electronegativity | 0.00000 | 0 | 1.00 |
| 9 | following | 0.00000 | 0 | 1.00 |
| 10 | heat | 0.00000 | 0 | 1.00 |
| 11 | highest | 0.00000 | 0 | 0.00 |
| 12 | ice | 0.13899 | 0 | 1.00 |
| 13 | ice | 0.27799 | 0 | 1.00 |
| 14 | increase | 0.00000 | 0 | 1.00 |
| 15 | kerosene | 0.00000 | 0 | 1.00 |
| 16 | kerosene | 0.27799 | 0 | 1.00 |
| 17 | level | 0.00000 | 0 | 1.00 |
| 18 | man | 0.00000 | 0 | 1.00 |
| 19 | measure | 0.00000 | 0 | 1.00 |
| 20 | melts | 0.00000 | 0 | 0.00 |

CNN/SVM Outcomes

| instanceId | predvalue |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |
| 8 | 0 |
| 9 | 0 |
| 10 | 0 |
| 11 | 0 |
| 12 | 1 |
| 13 | 0 |
| 14 | 0 |
| 15 | 1 |
| 16 | 0 |
| 17 | 0 |
| 18 | 0 |
| 19 | 0 |

Tags Predicted

ice,kerosene,Young`s,8,9,animal,banana,color,comes,creature,does

**Figure 2:** Evaluating keywords based on their weights

Figure 2 shows the evaluation of CNN based Automatic question tagging system where the keywords are tabulated based on their weights. Each term has its term frequency, position and an NLP Score. By analyzing all of these factors the tags are predicted.

Note that, all the questions are needed to be in multiple choice formats with a question followed by a set of 4 options.

## 4. RESEARCH OBJECTIVE & METHODOLOGY

### *A. Objective*

The paper deals with text mining, where tags are predicted based on the Convolution Neural Network. Within each convolution layer the keywords are filtered out and generate the resultant tag. Tagging can also be used to identify the subject based on the given question. NLP based automatic tagging only helps to tag questions based on English words. In Natural Language Processing, LDA summarizes the content of a document quickly by seeing the set of words it uses. LDA uses topic modelling to analyse the questions and develops tags based on the knowledge map.

### *B. Advantage of this system*

The automatic question tagging system tags the questions which can be helpful for many online examinations like Computerized Adaptive Testing (CAT). These types of questions need tags which can be used for students to help solve it.

Automatic tagging helps to identify the questions based on their subject. The tags are the mined product which is produced after learning questions.

NLP is used for predicting the tags based on the English words from the given corpus. But LDA can be used to give tags which is more topic related. These tags will produce a more accurate result.

Tagging are processed in a Convolutional Neural Network where each layer filter the neurons based on the weights and biases.

### *C. Methodology*

ConvNet or Convolutional Neural Network contains a list of layers which transforms the input volume into an output volume. Each of the layer accepts an input 3D volume and converts it to an output 3D volume through a differential function. The basic working of a Conv Layer follows the

The input volume is filtered out in each layer. The input volume is structured as 3D volume space and each of the filter applies a dot product to produce a next layered output. This process continues until a required output layer is obtained.

LDA is a probabilistic model for discovering latent semantic topics in large collection of text data. A random mixture of topics that mentions the proportion of time the document spends on each topic then categorizes each financial document.

*D. Experiment Result*

Using Automatic Tagging system, some experiments are conducted based on accuracy, precision and time complexity. A multiple choice question paper is selected for this experiment. The observations are taken based on the accuracy, precision and time complexity and drawn into graph. Figure 3 represents the graph which shows the tagging.

The graph shows the question paper being tested using both CNN and CNN using LDA. The graph varies for each of the trial numbers.

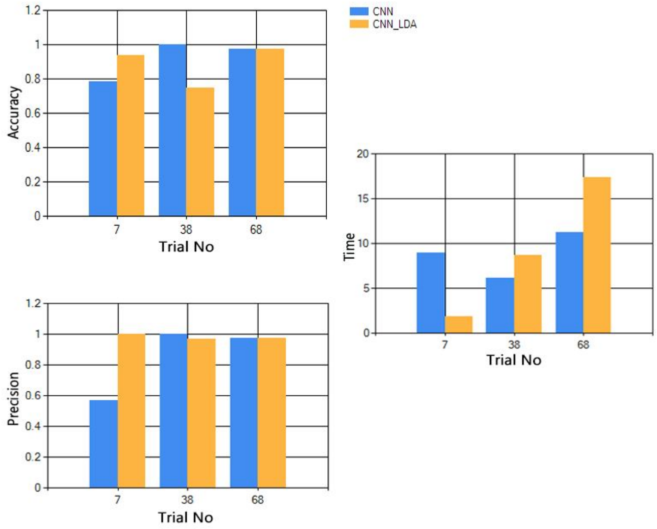


**Figure 3:** Graph showing precision, accuracy and time complexity using CNN and CNN using LDA

The graph changes with question paper because evaluation using both methods is different and all the 3 factors may vary.

## 5. FUTURE WORK

Automatic question tagging system can be extended to predict the import words within document and how they are related to other documents within a corpus. The semantic LDA can have a wide range of possibilities in topic modelling, which includes image prediction. Handwritten words, and extracting and generating a words within a document. This system can further developed to identify where the question resides in document and can include a graph which can mention subject weightage of the question bank.

## 6. CONCLUSION

CNN based automatic question tagging system is used to generate tags which are predicted from the given question. Each of the words within the documents are assigned with weights and which are further used in the convolutional layers. The tags can be generated with NLP and LDA. LDA is more accurate than NLP as it mines deeper into the topics rather on English words. Automatic question tagging opens an opportunity to tag questions which would be analyses the way of solving it.

## REFERENCES

1. Bo Sun, Yunzong Zhu, Yongkang Xiao, Rong Xiao and Yungang Wei, **Automatic Question Tagging with Deep Neural Networks**, *IEEE Trans. on Learning Technologies*, 2019.
2. SonamSonam, AyushiVermaSangeetaLal, NeetuSardana,. **TagStack: Automated System for Predicting Tags in StackOverflow***International Conf. on Signal Processing and Communication*, 2019.
3. InbalYahav, Onn Shehory, and David Schwartz.**Comments Mining With TF-IDF: The Inherent Bias and Its Removal**,*IEEE Trans. Knowledge and Data Engineering.*
4. Zhang Kunpeng, **Research on the Optimizing Method of Question Answering System in Natural Language Processing**, *International Conf. on Virtual Reality and Intelligent Systems*. 2019.
5. Qingqiang Wu, Xiang Deng, Caidong Zhang, Changlong Jiang, **LDA-based Model for Topic Evolution Mining on Text**, *The 6th International Conf. on Computer Science & Education.*
6. Ms. Pooja Kherwa1, Dr.Poonam Bansal**, Latent Semantic Analysis: An Approach to Understand Semantic of Text,** *International Conf. on Current Trends in Computer, Electrical, Electronics and Communication.* 2019.
7. Taniya Sasini, Sachin Tripathi, **Predicting tags for Stack Overflow Questions Using Different Classifiers,** *4th Int'l Conf. on Recent Advances in Information Technology*.
8. Liu Lin, Fan Xiaozhong, Zhao Xunping, **Research on Web Monitoring System Based on Natural Language Processing,** *International Conference on Natural Language Processing and Knowledge Engineering.*
9. Tamanna Siddiqui, Abdullah Yahya Abdullah Amer, Najeeb Ahmad Khan, **Criminal Activity Detection in Social Network by Text Mining: Comprehensive Analysis**, *4th International Conference on Information Sysyemms and Computer Networks (ISCON),* 2019.
10. MagantiSyamala, N.J. Nalini**, LDA and Deep Learning: A Combined Approach for Feature Extraction and Sentiment Analysis,** *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019.*
11. NontKanungsukkasem andTeerapongLeelanupang**, Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction,** *International Conference on current trends in Computer.*