International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 1–5

Generation of Question and Answer from Unstructured Document using Gaussian Mixture Neural Topic Model

[1]G.Divyabharathi, [2]Dr.G.Anupriya
[1]PG Scholar, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[1]divyacse02@gmail.com, [2]anuraj@drmcet.ac.in

Abstract: *Question Answering (QA) system is one of the ever growing applications in Natural Language Processing. The purpose of Automatic Question and Answer Generation system is to generate all possible questions and its relevant answers from a given unstructured document. Complex sentences are simplified to make question generation easier. The accuracy of the generated questions is measured by identifying the subtopics from the text using Gaussian Mixture Neural Topic Model (GMNTM).The similarity between generated questions and text are calculated using Extended String Subsequence Kernel (ESSK). The syntactic correctness of the questions is measured by Syntactic Tree Kernel which computes the similarity scores between each sentence in the given context and generated questions. Based on the similarity score, questions are ranked. The answers for the generated questions are extracted using Pattern Matching Approach. This system is expected to produce better accuracy when compared with the system using Latent Dirichlet Allocation (LDA) for subtopic identification.*

Keyword: *Privacy; k-anonymity; l-diversity; MSB-KACA, t-closeness; Particle Swarm Optimization;*

1. INTRODUCTION

Artificial Intelligence is one of the emerging technologies in recent days. It plays a major role in making computers to be intelligent. Natural Language processing is an important area in Artificial Intelligence. NLP understands human languages and acts accordingly. Question Generation and Answering is an application of NLP, where the objective is to generate set of questions from a collection of documents and to find the exact answer for the questions. The QA system receives a question as input and best, possible correct answers as output [1].

In this work, the questions are automatically generated from the topics and the answers are also extracted for the generated questions. It is assumed that each text will have some useful information that falls under a topic. Fact based questions about a given topic are generated from the contentbased on the named entity information and predicate argument structures of the sentences. During question generation, the relevancy of the questions to the topic must be considered as an important factor. The specific meaning in which the word is used in the sentence would help to improve the relevancy of the generated questions. Semantic role labeling [2] is used to identify the roles from the sentences to generate questions related to the content.

To ensure whether the question is closely related to the topic and speaks about the topic, Topic Relevancehas been measured by identifying subtopics in the content. Technique named Latent Dirichlet Allocation (LDA) has beenused for subtopic identification [3]. Extended String Subsequence Kernel (ESSK) has been applied to calculate the similarity of the questions. Since LDA fails to consider the order of the words and its semantics, it is proposed to use GMNTM model [4]. Syntactic correctness of the generated questions can be measured and ranked based on similarity scores. By using a pattern matching technique, the answers for the generated questions can be extracted.

The rest of this paper is organized as follows. Section 2 reviews related work and Section 3 discusses on methodology. Section 4 discusses about the experimental results and Section 5 presents the conclusion.

2. LITERATURE SURVEY

Question Answering (QA) system is concerned with building systems that automatically answer questions posed by human natural language. But generating questions is tedious as the questions may or may not exactly relate to what actually the document means. Various researchers proposed different approaches for efficient question and answer generation [5].

Heilman et al [6] have discussed the semantic issues in complex sentences. They provided an extraction algorithm for simplifying the complex sentences by removing appositives, clauses etc. The extraction process has two steps. (1) Extracting simple sentences from a given text (2) Moving phrases and quotes to the end of the verb phrase.

Chui et al [7] proposed information abstraction techniques like anaphora resolution and factual statement extraction. Factual statement extractor is used for generating questions. The abstractive technique used by this extractor seems to be the better technique when compared to others for sentence simplification.

Automatic question generation system based on semantics was proposed by Fattoh et al [8]. It is based on semantics using both semantic role labeling (SRL) and named entity recognition (NER) technique. The questions are generated based on extraction of attributes from NER and SRL.

Latent Dirichlet Allocation (LDA) has been used for identifying topics in the text document. Blei et al [3] proposed a model where each document is generated as a mixture of topics and continuous valued mixture proportions are distributed as Latent Dirichlet Random Variable. LDA topic modeling has been used for finding best result in topic identification but LDA works on bag-of-words assumption which does not bother about the order of words occurring in the sentences.

Yang et al [4] came up with Gaussian Mixture Neural Topic Model (GMNTM) as an alternative to LDA. It incorporates both the ordering of words and the semantic meaning of sentences into topic modeling. Each topic is represented as a cluster of vectors and embeds the corpus into a collection of vectors generated by the Gaussian Mixture Model. The surrounding words for a word are also considered to learn better topics and more accurate word distributions for each topic. This model has significantly better performance in terms of retrieval accuracy and classification accuracy.

Extended String Subsequence Kernel (ESSK) has been used for similarity measure calculation and has been proposed by Sadid et al [9]. WordNet has been used for solving Word Sense Disambiguation. The senses from the Disambiguation graph were taken and compared with subtopics leading to similarity score computation. By comparing with various measures like BOW, N-gram, TREE and WSK, Suzuki et al. [10] has found that ESSK produces better results.

Though various approaches are used in Question Generation and Answering (QA) systems, LDA is used for subtopic identification. But LDA treats the text as bag-of-words and fails to consider its order and semantics. To overcome this drawback of LDA, a model called GMNTM is used. It includes both the ordering of words and the semantic meaning of sentences into topic modeling.

3. METHODOLOGY

The system has been designed to generate all possible questions from content and also find answers for the generated questions. The system works mainly in five steps. The overall system architecture is given below in Figure 1.

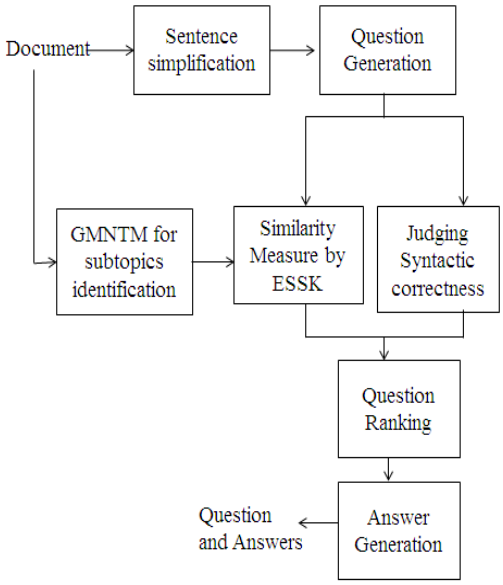

Figure 1 System Architecture

In the first step, complex sentences from the body of texts are simplified, to reduce the complexity of generating questions. In the second step, generic questions are generated from the extracted named entity information and the predicate argument structures. In the third step, subtopics are identified from the body of text by using LDA. ESSK is applied to find the similarity between the questions generated previously and the subtopics identified by LDA. Since, LDA does not bother about order of the word and its semantics, GMNTM model is used. In step four, syntactic

similarity between the generated questions and the sentences in the text document is measured by syntactic tree kernel. This helps to generate questions with correct syntactic structure. The generated questions are then ranked by using the ESSK similarity scores and the syntactic similarity scores. In the final step, answers for the generated questions are extracted by pattern matching technique.

3.1 Sentence Simplification and Named Entity Information

Sentences might have numerous clauses and complex grammatical structure. Hence the complex sentences are simplified to make question generation easier and accurate. By using factual statement extractor model [6], complex sentences are simplified. This model extracts the simpler form of the complex sentence. It modifies syntactic structure and semantics, lexical items along with phrase types such as leading conjunctions, appositives for generating the simplified sentences and sentence-level modifying phrases. The simple sentences are processed to generate all possible questions from the content. Initially, the simple sentences are passed to a Named Entity Tagger. The plain text is tagged into named entities. After tagging, general purpose rules are used to generate basic questions without bothering about the presence of answers in the body of text. This is to produce variety of questions and helps to answer even if there is no knowledge on a particular topic.

3.2 Semantic Role Labeling (SRL)

Semantic role labeling is used to generate specific questions from the simplified sentences. The sentences are semantically parsed using SRL system. An automatic statistical semantic role tagger is used to annotate naturally happening text with semantic arguments. The syntactic analysis is done fully when a sentence is provided to the role tagger. All the verb predicates are identified and features are extracted for all constituents in the parse tree comparative to the predicate. Finally the constituents are tagged with the suitable semantic arguments. The semantically labeled sentences are transformed into questions using general purpose rules [11].

3.3 Sub Topic Identification

To ensure the generated questions are correct or relevant to the topic, Sub topic identification from the body of text is necessary. Topic Relevance measurement is a common way to check whether the question is asking something about the topic or something that is very closely related to the topic. This identifies the subtopics in the given body of the text. In [3], Latent Dirichlet Allocation (LDA) is used for subtopic identification. LDA is a probabilistic topic modeling technique. LDA considers each document as different topics. It works on the assumption that all documents are a bag-of-words [12] and selects a distribution over topics. For each word in the document, a topic is chosen randomly agreeing to the distribution and a word is drawn from the topic.

$$P(w_i)=\sum_{j=1}^{k} P(wi|zi = j)P(zi = j)$$

Where k denotes the number of topics, $P(w_i|z_i=j)$stands for the probability of word w_i under topic j, and $P(z_i=j)$ is the sampling probability of topic j for the ith word.

3.4 GMNTM Model

Even LDA works on the assumption that the documents are simply a bag-of-words. This assumption does not bother about the order of words and also do not properly capture the exact semantics of the context. Gaussian Mixture Neural Topic Model represents the topic model as Gaussian mixture model of vectors. This encodes words, sentences, documents. Mixture component is associated with a specific topic individually. This method learns the topic model and the vector representation jointly. The position of the word is learnt to optimize the predictability of the word using the words surrounding it provided that the vector representations are sampled from the Gaussian mixtures which represents the topics. The main idea is the semantic meaning of the sentences and documents are fused to infer the topic of a specific word [4].

The generative model of GMNTM is an assumption that there are W different words in the vocabulary where each word w ϵ {1…, W} along with its vector representation. Similarly the documents in the corpus and the sentences in the documents are indexed and also have their own vector representations. There are T topics in the GMNTM model where T is selected by the user. Each topic corresponds to a Gaussian mixture component. V-dimensional Gaussian distribution $N(\mu_k, \sum_k)$ with mixture weight $\Pi_k \epsilon$ R. The parameters of Gaussian mixture model are collectively represented by

$$\lambda = \{\Pi_k, \mu_k, \sum_k\} \quad k=1,\ldots,T \quad (1)$$

Given the Gaussian mixture model λ the generative process model is stated below, for each word w in the vocabulary its topic is sampled along with its vector representation from its Gaussian distribution. Simultaneously the vector vec(w) is also sampled. For each document d the sentences in the document are

sampled with its vectors. The entire collection of latent vectors is denoted by ψ.

$\Psi := \{vec(w)\} \cup \{vec(d)\} \cup \{vec(s)\}$ (2)

For each word slot in the sentence, the realization of the word is generated according to the documents vector and current sentence vector as well as most m prior words in the same sentence. The word realization is represented by w_i and is defined as,

$P(w_i = w|d, s, w_{i-m}, \ldots . w_{i-1})$

$\alpha \exp(a^w_{doc} + a^w_{sen} + \sum_{t=1}^{m} a^w_t + b)$

Where a_{doc},a_{sen} and a_t are influences from the document d, the sentence s and the previous word w, respectively. They are defined by

$a^w_{doc} = <(u^w_{doc}, vec(d))>$

$a^w_{sen} = <u^w_{sen}, vec(s)>$

$a^w_t = <u^w_t, vec(w_{i-t})>$

Here, $u^w_{doc}, u^w_{sen}, u^w_t \in R^V$ are parameters of the model that are shared across all slots in the corpus. U is to represent this collection of vectors,

$U := \{u_{doc}, u_{sen}\} \cup \{u_t \mid t \in 1,2, \ldots \ldots, m\}$

Estimating all these model parameters the word representations are learnt. This method has been found to be very effective for identifying topics [4].

3.5 Similarity Measure

Extended String Subsequence Kernel (ESSK) is a technique used to measure the similarity [10] between generated questions and subtopics identified. Each word is considered as an alphabet and the alternate is all possible sentences. A Dictionary-based Word Sense Disambiguation (WSD) system [13] is used for assuming one sense per discourse. WordNet[14] is used to find the semantic relations for the words in the text. The semantic relations are assigned with weights based on heuristics. The WSD technique is decomposed into two steps: (1) All possible senses of the words are constructed as representation and (2) Based on the highest score, the words are disambiguated. Each word from the text is expanded to all of its possible senses. A disambiguation graph is constructed with nodes and edges, where the nodes denote the senses and the edges denotes semantic relations. This graph is exploited to perform the WSD. The weights of all edges are summed, leaving the nodes with different senses. The sense with highest score is considered as the most probable sense. In case of a tie between two or more senses, the sense which comes first in WordNet is selected.

ESSK is used to measure the similarity between the generated question word/senses and topic word/senses. Similarity score Sim(T_i,Q_j) is calculated using ESSK, where T_i is a topic/sub-topic and Q_i denotes the generated question. ESSK is defined as follows:

$$K_{essk(T,Q)} = \sum_{m=1}^{d} \sum_{t_i \epsilon T} \sum_{q_j \epsilon Q} (t_i, q_j)$$

$$K_m(t_i, q_j) = \begin{Bmatrix} val(t_i, q_i) if m = 1 \\ K'_{m-1}(t_i, q_j).val(t_i, q_j) \end{Bmatrix}$$

where, t_i and q_j are nodes of T and Q, respectively and d refer to the vector space dimension. The function val(t,q) returns the number of common attributes (i.e., the number of common words/senses) to the given nodes tand q.

$$K'_m(t_i, q_j) = \begin{Bmatrix} 0 & if j = 1 \\ \lambda K'_m(t_i, q_{j-1}) + K''_m(t_i, q_{j-1}) \end{Bmatrix}$$

where, λ is the decay parameter for the number of skipped words. $K''_m(t_i,q_j)$ is defined as follows,

$$K''_m(t_i, q_j) = \begin{Bmatrix} 0 & if i = 1 \\ \lambda k''_m(t_{i-1}, q_j) + K_m(t_{i-1}, q_j) \end{Bmatrix}$$

After normalization the similarity measure is defined as follows,

$$sim_{essk}(T, Q) = \frac{K_{esk}(T, Q)}{\sqrt{K_{essk}(T, T) K_{essk}(Q, Q)}}$$

3.6 Syntactic Correctness and Ranking

The next step of the system is to determine the syntactic correctness of the generated questions. In order to reduce the human intervention to check the syntactically incorrect questions generated, the tree kernel functions are applied and re-implement the syntactic tree kernel model. The sentences and questions are parsed into syntactic tree using Charniakparser [15]. Similarity between two equivalent trees is measured using tree kernel method [16]. This kernel function calculates the number of common sub trees between the two. Based on the syntactic structure, it produces similarity scores of sentences in the text and generated questions.The average of similarity scores are computed which is used for ranking the generated questions.

3.7 Answer Extraction

The Pattern Matching Approach uses a bootstrapping approach called Snowball technique to collect the answer model [17]. It builds the answer model for a class of questions to capture all kind of answer patterns and the similar answer for the generated questions can be obtained. Snowball technique uses collection of pairs in the document to find all occurrences of pairs and these pairs are further used for pattern matching. In pattern matching, the patterns are matched with the questions and the matched patterns are extracted as the exact answers for the generated questions.

4. RESULTS

TABLE I SUBTOPICS IDENTIFICATION BY LDA

Topics	Score	Topics	Score
Violin	8	Music	6
Bowed	7	String	8
Instrument	2	Tuned	8
Perfect	2	Smallest	6
Highest-pitched	4	Informally	3

Question : What does the word ``violin'' mean ?
Topic Similarity : 55.31109002304385
Normalization : 7.82216936602278
Question : What comes from the Middle Latin word virtual?
Topic Similarity : 38.438209691539406
Normalization : 6.691233442236767
Question : Who is called a violinist or a fiddler?
Topic Similarity : 125.24032378357596
Normalization : 11.887285470981299
Question : Who is a person who makes or repairs violins called?
Topic Similarity : 124.57663948312056
Normalization : 11.46820922912309

Figure 2 Similarity score calculation by ESSK

The similarity score calculation between the topics generated by LDA and the questions are shown in Figure 2. Topic relevance and syntactic correctness are the two performance evaluation metrics from which the accuracy is obtained. The system accuracy is yet to be measured and replacing LDA with GMNTM model is expected to improve the performance of question and answer generation system.

5. CONCLUSION

Question answering system needs to retrieve specific information from the text rather than the whole documents. Finding the questions from given documents could be made easier by sub topics identification. To overcome the drawbacks of LDA, a topic model called GMNTM is applied instead of that. This variation in the system might produce better results. As an enhancement answer generation is also included along with questions. Pattern matching approach is chosen as a solution for accurate answer generation. The combination of GMNTM and Pattern matching approach might be a good solution for automatic question answer generation system.

REFERENCES

[1] Cooper, R.J., and Ruger S.M.,(2000), “A Simple Question Answering System”, In Text Retrieval Conference(TREC).
[2] Chali, Y., and Sadid A. Hasan., (2015), "Towards Topic-to-Question Generation" Computational Linguistics, Vol.41, pp. 386-397.
[3] Blei,D. M., Ng, A. Y., and Jordan, M. I., (2003), “Latent Dirichlet allocation”, Journal of Machine Learning Research, Vol. 3, pp.993–1022.
[4] Yang, M., Cui, T., and Tu, W.,(2015), “Ordering-Sensitive and Semantic-aware Topic Modeling”,Twenty-Ninth AAAI conference on Artificial Intelligence.
[5] Kim, S.M.,Baek, D.H., Kim, S.B., and Rim, H.C., (2000), “Question Answering Considering Semantic Categories and Co-occurerence Density”, In Text Retrieval Conference(TREC).
[6] Heilman, M., and Smith, N.A., (2010), “Extracting simplified statements for factual question generation”, The Third Workshop on Question Generation, Pittsburgh, pp.11-20.
[7] Chua, D.D., Aqunio, J.F., Kabiling, R.K., Pingco, J.N., and Sagum, R., (2011), “Text2 Test: Question Generator utilizing information abstraction techniques and question generation methods for narrating and declarative text”, Proceedings of the 8th National Natural Language Processing Research Symposium, Manila, pp. 29-34.
[8] Fattoh, I.E., Aboutabl, A.E., and Haggag, M.H., (2014), “Semantic Based Automatic Question Generation using Artificial Immune System”, Computer Engineering and Intelligent Systems, Vol.5, pp. 74-82.
[9] Chali, Y., Sadid, A., Hasan and Imam, K., (2011), “Using semantic information to answer complex questions”, Computational Linguistics, Vol.4, pp. 68-73.
[10] Hirao, T., Suzuki. J, Isozaki. H, and Maeda. E., (2004), “Dependency-based sentence alignment for multiple document summarization”, In Proceedings of COLING (International conference on Computational linguistics), Geneva, pp. 446–452.
[11] Kingsbury, P., and Palmer, M., (2002), “From Treebank to PropBank”, In Proceedings of the International Conference on Language Resources and Evaluation, Las Palmas,pp.1, 989-1,993.
[12] Misra, H., Cappe, O., and Yvon, F., (2008), “Using LDA to detect semantically incoherent documents”,In Proceedings of the Twelfth Conference on Computational Natural Language Learning, Manchester, pp. 41-48.
[13] Chali, Y. and Joty. S. R., (2007), “Word sense disambiguation using lexical cohesion”, In Proceedings of the 4th International Conference on semantic Evaluations, Prague, pp. 476–479.
[14] Fellbaum, C., (1998), “WordNet – An Electronic Lexical Database”, MIT Press.
[15] Charniak, E.,(1999), “A maximum-entropy-inspired parser”, Technical Report CS-99-12, Brown University, Computer Science Department, Rhode Island.
[16] Collins, M., and Duffy, N., (2001), “Convolution Kernels for Natural Language”, In Advances in Neural Information Processing Systems, Vancouver, pp. 625-632.
[17] Agichtein,E.,andGravano,L.,(2000),”Snowball:Extracting Relations from Large Plain-text Collection”,Proceedings of the fifth ACM conference on Digital Libraries, ACM, pp.85-94.