Answer Extraction System Based on Latent Dirichlet Allocation

Mohammed A. S. Ali[1], Sherif M. Abdou[2]
Information Technology Department
Faculty of Computers and Iinformation,Cairo University
Cairo, Egypt

***Abstract*—Question Answering (QA) task is still an active area of research in information retrieval. A variety of methods which have been proposed in the literature during the last few decades to solve this task have achieved mixed success. However, such methods developed in the Arabic language are scarce and do not have a good performance record. This is due to the challenges of Arabic language. QA based on Frequently Asked Questions is an important branch of QA in which a question is answered based on pre-answered ones. In this paper, the aim is to build a question answering system that responds to a user inquiry based on pre-answered questions. The proposed approach is based on Latent Dirichlet Allocation. Firstly, the dataset, pairs of questions and associated answers, will be grouped into several clusters of related documents. Next, when a new question to be answered is posed to the system, it,therefore, starts to assign this question to its appropriate cluster, then, use a similarity measure to get the top ten closest possible answers. Preliminary results show that the proposed method is achieving a good level of performance.**

***Keywords*—Question Answering; frequently asked questions; information retrieval; artificial intelligence;**

I. Introduction

Question Answering (QA) is a task which has been created to satisfy the specific and urgent need of a user to get a direct answer to a given question. Generally speaking, QA tasks can be classified, from an information retrieval perspective, into two separated types: question answering based on retrieving and forming an answer from flat documents, and the second one is based on retrieving an answer to a similar pre-answered question. Both types are considered active research topics in information retrieval. However, this paper is concerned only with QA based on the Frequently Asked Questions (FAQ) task.

According to the literature, great efforts have been made to build a reliable QA system. However, few of these attempts have been made for the Arabic language. And among those only a few of them are oriented to QA based on a FAQ task. This lack of such systems is due to challenges presented by the Arabic language.

Arabic is a Semitic language spoken as a native language by more than 330 million people [1]. Arabic is a morphologically complex, highly derivational and inflectional language. Moreover, Arabic is rich in the use of affixes and clitics and, usually, disambiguating short vowels and other orthographic diacritics in standard orthography are omitted [2]. Therefore, it has been difficult, to some extent, to build a reliable QA system.

In this paper, a system for QA based on Latent Dirichlet Allocation (LDA) [3] has been presented. The LDA has been exploited, as a clustering algorithm, to divide the dataset into related document groups. Then, its estimated models parameters has been also exploited to calculate the similarity between the new question and each question-answer pair in its closest group [4].

The domain in which this application will be applied is Islamic Fatwa. A Fatwa is a formal Islamic legal opinion issued by expert scholar(s) (mufti or committee) in response to a question from an individual. In Fatwa, mufti clarifies an issue based on evidence from Shariah [5]. The Fatwa is considered as an Islamic religion verdict, therefore, Muslims all over the globe are interested in them and seek them out on a daily basis. Moreover, the field is very sensitive, so, mistakes are not allowed. The official Fatwa organizations are responsible for receiving, handling and replaying these daily questions.

Due to the limitation of human resources, these organizations are unable to handle this barrage of questions within a reasonable time frame. Meanwhile, many newly posed questions have similar answered ones in the database. Unfortunately, there are no effective and reliable systems yet built to automatically retrieve such a type of questions.

A. Previous work

Several pieces of research have been proposed in the literature in the field of question answering based on already pre-answered ones.

In [6] R. D. Buke el.al. have proposed a system to fetch a similar question to a newly posed question. This system was called FAQFinder. Their system is based on a vector space model, and included a semantic definition of similarity between words based on the concept of hierarchy in WordNet as well.

Keliang Jia et.al. in [7] have built a QA system for remote learning applications, so as to enhance the communication facilities between teachers and their students. They calculated the similarity between questions by integrating both similarity between domain keywords using a domain knowledge dictionary and similarity between common words using HowNet.

Zhiguo Wang et.al in [8] have tried to address the issue of FAQ-based QA via word alignment. They started with extracting a feature vector, including (similarity, dispersion, penalty, 5 important words, reverse and some spare lexical features), from pair (query, candidate), then used a neural network to calculate the similarity between such a pair.

None of the previously mentioned works is concerned with Arabic language. However, in [9], [10] Islam Elhelwany et.al. have proposed an Arabic Fatwa Intelligent system based on textual case based rezoning which was firstly used in [11]. In their system, they started by extracting a representative term for each cluster which were later called clusters attractors. Then, the cases clustered around these attractors. Eventually, they used Jensen-Shannon divergence to assign a newly posed question to its appropriate cluster and, subsequently, to find the closest possible question among questions in such a cluster. Unfortunately, no results or evaluation are presented in these works and the data sets are not available for comparison. In general, none of the existing works efficiently addresses the task of Arabic QA based on FAQ which is going to be address in this work.

The rest of this paper is organized as follows: Section 2 introduces the proposed method; the evaluation and experimental results are discussed in Section 3; and finally, in Section 4, our findings are summarized and some future work is propose.

II. Approach

A. Latent Dirichlet Allocation (LDA) model estimation

Latent Dirichlet Allocation (LDA) is an unsupervised, statistical approach for document modeling that discovers latent topics in a collections of text documents, in this case each document is Fatwa (question and answer). LDA considers a word as a basic unit of information, and it assumes that documents that discuss similar topics use a similar collection of words. In other words, documents are modeled as distribution of topics (θ), and each topic is modeled as a distribution of words (ϕ). topics are thus discovered by recognizing collections of words which frequently occur together within documents [3]. In figure 1 a graphical representation of LDA is shown. As depicted in the figure [2] M is the number of documents of arbitrary length in the collection, T topics and V words forming the vocabulary. Here, the topic distribution per document and the per-topic word distributions are sampled from $Dir(\alpha)$ and $Dir(\beta)$ respectively. The LDA model estimation goes through these steps:

1) Choose number of topics T and LDA hyperparamters α and β.
2) For each document
    a) Choose the number of words N.
    b) For each word:
        i) Sample z from $\theta^{(j)}$, where j is the index of the current document.
        ii) Sample w from $\phi^{(z)}$

In LDA the goal is to estimate the distribution $p(z/w)$. Unfortunately, exact estimation of LDA parameters is an intractable problem. The solution to this problem is to use an approximation estimation algorithm; common methods to do so include Expectation propagation and Gibbs sampling [12] , which is more common and is followed here.

We will present only the most important equation used by the algorithm for topic sampling for words. Let $\vec{w}$ and $\vec{z}$ be the vectors of all words and their topic assignment of the whole documents collection W respectively. The topic assignment for a particular word depends on the current topic assignment of all the other word positions. More specifically, the topic assignment of a particular word t is sampled from the following distribution:

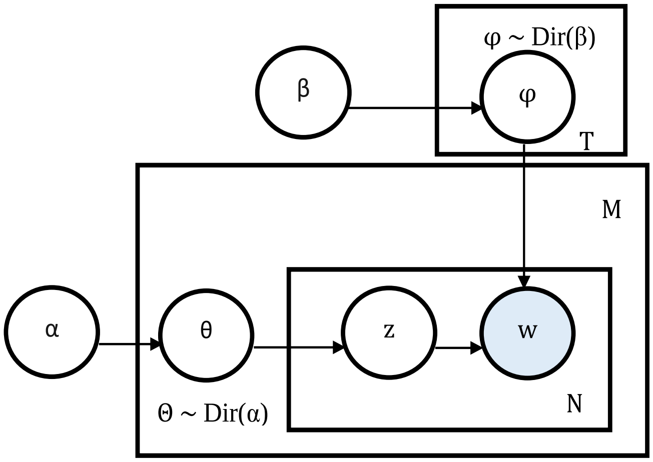

Fig. 1: LDAs graphical representation [3] shaded nodes represent observed variables whereas other nodes represent latent ones.

$$p(z_i = k \mid \vec{z_{\neg i}}, \vec{w}) = \frac{n_{k,\neg i}^{(t)} + \beta_t}{[\sum_{v=1}^{V} n_k^{(v)} + \beta_v] - 1} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{j=1}^{K} n_m^{(j)} + \alpha_j] - 1}$$

where $n_{k,\neg i}^{(t)}$ is the number of times the word t is assigned to topic k except the current assignment; $\sum_{v=1}^{V} n_k^{(v)} - 1$ is the total number of words assigned to topic k except the current assignment;$n_{m,\neg i}^{(k)}$ is the number of words in document m assigned to topic k except the current assignment; and $\sum_{j=1}^{K} n_m^{(j)} - 1$ is the total number of words in document m except the current word t. Here, α and β are hyperparamters of LDA and they describe the nature of the priors of θ and ϕ respectively.The choice of priors has an important implication for the result. For instance, choose high value for β can be expected to decrease the number of topics, whereas smaller β values will generate more topics [13].

Once p(z/w) is estimated using a sufficient number of Gibbs sampling iteration, the distributions ϕ and θ can be easily estimated using the following formulas:

$$\phi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v}$$
$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_m^{(j)} + \alpha_j}$$

once the LDA model is estimated, that is to say that the corpora have been clustered and the estimated model can be used to infer a new document as well.

B. Semantic answer extraction using LDA estimated model

A similarity between two documents d1 and d2 is computed by multiplying both, the similarities between the topic distribution per-document ($\theta\ d1$ and $\theta\ d2$) and the per-topic word distributions ($\phi\ t1$ and $\phi\ t2$) [4]

The similarities between the topic distribution per-document can be calculated using the following equations:

$$IR(p,q) = \sum_{i=1}^{T} p_i\ log\frac{2xp_i}{p_i+q_i} + \sum_{i=1}^{T} q_i\ log\frac{2xq_i}{p_i+q_i} \tag{1}$$

such that p and q are the document distribution over topics for $d1$ and $d2$ respectively.

As IR measures the distance between two documents, it can be transformed into similarity measure using the following equation:

$$SIM(p,q) = 10^{-\delta IR(p,q)} \tag{2}$$

Meanwhile, each word has a certain contribution to a topic (word distributions per-topic $\phi\ t1$ and $\phi\ t2$). Based on these contributions, word-to-word semantic similarity is defined. The word-to-word semantic similarity measure based on LDA is further used in conjunction with an optimal matching method to measure similarity given two documents. The similarities between word distributions per-topic can be considered as an assignment problem. Given a complete bipartite graph, $G = (D1, D2, E)$, with n document 1 vertices (words) $(D1)$, n document 2 vertices (words) $(D2)$, and each edge $e_{d1\in D1, d2\in D2} \in E$ a non-negative weight (similarity between the two words). The aim to find matching M from $D1$ to $D2$ with maximum similarity. Such an assignment is called optimum assignment. Method in [14] is used to solve this assignment and can be formulated as finding a permutation π for which $q_{OPT} = \sum_{i=1}^{n} word\text{-}sim(d1_i, d2_{\pi(i)})$ is maximum such that $word\text{-}sim$ is word-to-word similarity measure based on LDA and can be calculated using the following equation (Hellinger distance)

$$HD(w1,w2) = \frac{1}{\sqrt{2}} + \sqrt{\sum_{1}^{T}(\sqrt{w1_i} - \sqrt{w2_i})^2} \tag{3}$$

Such that $w1$ and $w2$ are the word distributions per-topic for $d1$ and $d2$ respectively.

Briefly, the proposed method can be summarized in these steps:

1) Performed data pre-processing.
2) Estimate LDA model for the collection of documents.
3) Cluster the document collection based on estimated LDA model.
4) When a new question is posed, assign it to its appropriate cluster using LDA inferencer.
5) Once the new question is assigned to its cluster, retrieve the ten closest answers possible using measures mentioned in section II-B
6) Display results

III. Results and discussion

The dataset that has been used was collected from the well-known website that introduces Islamic Fatwas "IslamWeb"[1]. The total number of documents is 11109. Each one of these documents represents a Fatwa which contains a question and associated answer. All documents in this collection are used to estimate LDA model in step II-A described in the methodology. For the test, 110 non-answered questions were posed to the system and the result obtained shown to seven educated users. The users were then asked to tell how much they agreed with the following statement: "this answer fits my question and I am satisfied with it". The users rated their degree of agreement on a 5-point Likert scale where 1 indicates strong disagreement and 5 indicates strong agreement.

It should be noticed that all results, of the proposed method, presented in this section are based on the following parameters which have been experimentally set: the Dirichlet hyper-parameters α and β were chosen to be 0.5 and 0.1 respectively and the number of topics was chosen to be 100. Gibbs sampling is stopped after 1000 steps.

It is difficult to compare the proposed approach and the various approaches described in Section I-A because 'the software applications and the textual resources used in the experiments are unavailable'. Moreover, the results of the respective experiments are not conclusive.

For example, the work presented by Islam et.al [9] does not measure the effectiveness of the presented approach. Moreover, in another work presented by the same authors [10], the only results shown are clustering results. However, neither the software applications nor the textual resources used in the experiments are available for comparison.

The rest of the works are oriented to other languages but not Arabic. Therefore, the effectiveness of the proposed approach will be evaluated by comparison with Google search engine, where the top ten retrieved results are collected manually and compared to the top ten retrieved results by the proposed method as shown in section III.

To estimate the performance, the average Likert scale and average retrieval time are calculated. The average Likert scale is defined as follows: let U and Q be the total number of users and total number of questions respectively. $LS = \frac{1}{Q\times U}\sum_{n=1}^{Q}\sum_{m=1}^{U} S_{n,m}$ such that $S_{n,m}$ is a score given by a user n to a question m. To test the inter-rater reliability, the Kappa measure has been calculated. As shown in the table I, the performance of the proposed system is better. This success is mitigated, though, by the fact that the Google average response time is better than ours by orders of magnitude. Nonetheless, it is commonly assumed that Kappa values between 0.4 and 0.6 offer a moderate level of agreement, and therefore, both of them, LDA-Optimal and Google get a moderate agreement. In Figure 2 a diverging stacked bar chart shows the raw results based on user evaluation of the proposed system and of the Google search engine. It presents the Likert scale results of the criteria "this answer fits my question and I am satisfied with it". As it can be seen in the figure, that number of answers with which the users 'strongly agree' and 'agree' in the proposed approach is

[1] www.islamweb.net/fatwa/

TABLE I: average Likert scale, average response time and Kappa measures of LDA-Optimal and google

Method	Average 5-point Likert scale	Average retrieval time (Second)	Kappa
Google	2.65	**0.85**	**0.58**
LDA-Optimal	**3.75**	22.4	0.55

clearly greater than those found through Google. Meanwhile, according to user evaluation, more than half of the questions are not answered through Google.

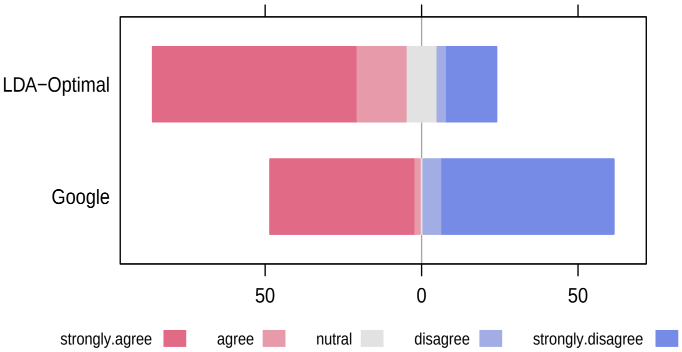

Fig. 2: Results of criteria this answer fits my question and I am satisfied with

After analysis of results for both methods, it has been found that, Google was able to handle those questions which contain only small number of words, 3-7 words, better than LDA-Optimal. On the other hand, the proposed approach has an ability to handle those questions with longer scripts, while Google has a lesser ability to do so and sometimes fails when the number of characters exceeds its limit.

IV. Conclusion and Future Work

With the boom of the Web's content, an inevitable need for an effective information retrieval system is required. In particular, the possibility of extracting a direct answer to a specific question. This process is called question answering and is currently one of the most active research areas in the field of information retrieval. The QA based on FAQ is the task in which a new question is answered based on already pre-answered ones.

In this paper, a new methodology is proposed to accomplish the task of QA based on FAQ. This approach assumes that an answer is a contextual expansion of its corresponding question. Therefore, the question and its associated answer is treated as one document. Since organization of documents into clusters of related documents has been shown to significantly improve the results of information retrieval systems, the approach first started to cluster the corpora into several clusters of related documents. Such clustering is achieved by the LDA model. When a new question to be answered is posed to the system, it is inferred, and assigned to an appropriate cluster using LDA inferencer.

Up to now, there is the question to be answered and its associated cluster. A similarity measure based on LDA estimated distributions is used to retrieve the closest possible answers to a given question.

In spite of all the advantages and possibilities of the proposed method, it has several limitations that could be improved in the future. First, the proposed approach does not consider the type of question, so future improvements to the accuracy of the system will involve a question analysis step so as to determine the type of question. Second, a different sophisticated similarity measure can be used instead of the current one. Finally, the current proposed approach does not handle negation, this may be dealt with in future researches.

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, p. 14, 2009.

[2] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 573–580.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] V. Rus, N. Niraula, and R. Banjade, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. Similarity Measures Based on Latent Dirichlet Allocation, pp. 459–470.

[5] J. Esposito and A. Sachedina, *The Islamic World: Hizbullah-Ottomon empire*, ser. The Islamic World. Oxford University Press, 2004. [Online]. Available: https://books.google.com.eg/books?id=GSUZAQAAIAAJ

[6] R. D. Burke, K. J. Hammond, V. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, "Question answering from frequently asked question files: Experiences with the faq finder system," *AI magazine*, vol. 18, no. 2, p. 57, 1997.

[7] K. Jia, X. Pang, and Z. Li, "Question answering system in network education based on faq," in *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*, Nov 2008, pp. 2577–2581.

[8] Z. Wang and A. Ittycheriah, "Faq-based question answering via word alignment," *CoRR*, vol. abs/1507.02628, 2015.

[9] I. Elhalwany, A. Mohammed, K. Wassif, and H. Hefny, "Using textual case-based reasoning in intelligent fatawa qa system," *International Arab Journal of Information Technology (IAJIT)*, vol. 12, no. 5, 2015.

[10] I. Elhalwany, A. Mohammed, K. T. Wassif, and H. A. Hefny, "Enhancements to knowledge discovery framework of {SOPHIA} textual case-based reasoning," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 211 – 220, 2014.

[11] D. Patterson, N. Rooney, M. Galushka, V. Dobrynin, and E. Smirnova, "Sophia-tcbr: A knowledge discovery framework for textual case-based reasoning," *Knowledge-Based Systems*, vol. 21, no. 5, pp. 404 – 414, 2008.

[12] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 91–100.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[14] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.