Intelligent Chatbot-LDA Recommender System

https://doi.org/10.3991/ijet.v15i20.15657

Yassine Benjelloun Touimi (✉), Adeladim Hadioui,
Noureddine El Faddouli, Samir Bennani
Mohammed V University, Rabat, Morocco
yassine.benjelloune.touimi@gmail.com

Abstract—With the proliferation of distance platforms, in particular that of an open access such as Massive Online Open Courses (MOOC), the learner finds himself overwhelmed with data which are not all efficient for his interest. Besides, the MOOC has tools that allow learners to seek information, express their ideas, and participate in discussions in an online forum. This tool is a huge repository of rich data, which continues to evolve, however its exploitation is fiddly in the search for information relevant to the learner. Similarly, the task of the tutor seems to be difficult in management of a large number of learners. To this end, the development of a Chatbot able to meet the requests of learners in a natural language is necessary to the deroulement a course in the MOOC. The ChatBot plays the role of assistant and guide for the learners and for the tutors. However, ChatBot responses come from a knowledge base, which must be relevant. Knowledge extraction to answer questions is a difficult task due to the number of MOOC participants. Learners' interactions with the MOOC platform generate massive information, particularly in discussion forums by seeking answers to their questions. Identifying and extracting knowledge from online forums requires collaborative interactions between learners. In this article we propose a new approach to answer learners' questions in a relevant and instantaneous way in a ChatBot in natural language. Our model is based on the LDA Bayesian statistical method, applied to threads posted in the forum and classifies them to provide the learner with a rich semantic response. These threads taken from the discussion forum in the form of knowledge will enrich the ChatBot knowledge database. In parallel, we will map the extracted knowledge to ontology, to provide the learner with pedagogical resources that will serve as learning support.

Keywords—Mooc, ChatBot, Forum Discussion , Latent Dirichlet Allocation, Knowledge extraction, ontology.

1 Introduction

The democratization of the use of the Internet and its penetration in the public and private space, made distance learning more economical and popular. As consequence, a mode of online learning has emerged in recent years and has attracted many followers, called Massive Online Open Courses (MOOCs) [8].

MOOCs differ from conventional online courses by having pre-recorded videos, open to an unlimited number of participants simultaneously for a single course, online quizzes, and integration of social networks in the learning curriculum. MOOCS are flexible in terms of learning style, and do not place requirements on learners. Access to the course requires only minimal information for the enrollments, and no cognitive requirements on the part of the learners [9]. However, this method of distributing information poses insurmountable challenges for tutors. The open access at any time and to any geographic area poses space-time constraints that require the omnipresence of the tutor.

In addition, the participation rate per course far exceeds the supervision ratio as in a traditional course, which can affect the engagement of online learners [11].In this case, a conversational agent can simulate human conversation in natural textual language with a virtual teacher to interact and answer learners' questions [10]. A conversational agent, or more commonly known as a chatbot, is a computer program designed to simulate a conversation with a human user. It is a virtual assistant that communicates with the user through text messages, usually in a messaging application (eg. Facebook Messenger, Telegram or Skype) or on websites [12].

The In a form of human communication, natural language [6] does not designate a language properly speaking, but the natural way of expressing itself in human beings, as opposed to binary language and the languages used in programming. It’s the language of emails, descriptions, chat, etc. Therefore, human language in its raw form cannot be exploited.

So, for the chatBot to understand what we want, it is necessary that the raw entry is processed to deduce the intention and keywords. Several techniques allow the creation of a ChatBot, and their main algorithms are accessible. They allow us to get an idea of the solutions we can develop. In addition, there are interesting and relatively easy to use tools.

However, they remain very limited and their main algorithms are inaccessible by way of example: Chatfuel, Rebot.me, Botsify, Wit.ai [13]. As well as a variety of languages have emerged to create ChatBots such as AIML [14], ChatterBot [15] and neural networks [16].

However, one of the biggest research challenges in developing effective chatbots is emulating human dialogues. Indeed, it turns out to be a difficult task and generates problems which are linked to the NLP (Natural language processing) research field [1]. Using NLP techniques and algorithms, it is possible to understand the purpose and request of the user.

Furthermore, it is difficult to map all user requests, because common ChatBots record a lack to predict the user's ideas during the conversation [2]. ChatBot technology is considered to be a way to remedy the pit that separates the human dimension from the machine, notably in terms of education [3]. ChatBot simulate the interactivity between the learner and the teacher, in a question-and-answer language, which plays an essential role in improving the skills of individual learning. There are many ways to use the chatbot in e-learning, such as monitoring of students by encouraging them to work by sending them notifications and reminders [6].

This aspect is evident in the experience of the participants in a MOOC, who claim the lack of interactions with the instructors, making the learning experience less satisfying and that the MOOCs are more suitable for curious learners who seek certifications more than 'consistent higher education [17].

Another point to note concerns learners who tend to multitask rather than focus on learning at their own pace, which takes them twice as long to complete. In addition, the learner is provided with the system for a personalized learning experience, so that each learner receives the information and acquires knowledge at their own pace, without being overloaded which can lead to abandonment [18].

In fact, this technology of bots perfects the interactions of the learners within the class and plays the role of moderator, guides or simulates the presence of the pairs of his community, and that of the teacher as if he works individually with the learner [19]. The ChatBot helps learners in their routine work, answering learners' questions and even checking their work.

Using ChatBots extends to the assessment of learners in real time, in the case of presence of several learners, it proves impossible to manage each alone, and becomes too tedious for the teacher. However, ChatBots can work with a multitude of students and groups at the same time [4].

Most existing ChatBots consist of dialog management modules to control the conversation process which use stimulus tags (question / answer), and manage knowledge bases with patterns to respond to user input manually [20].

Therefore, building ChatBots manually takes time and is difficult to adapt to new areas such as MOOCs. MOOCS are designed to be an environment for sharing ideas, knowledge, and discussing common topics, and promoting social interactions in a working group [21].

MOOCs have discussion forums [39] to supplement learning by interaction in the form of question / answer, or discussion. The data generated in the forums has been manipulated to retrieve information and to natural language processing. The mission of the discussion forum is to facilitate access to information rich in data, such as posts, questions, retrieval of responses, aggregation of threads.

On the other hand, discussion forums can cover the tagging of dialogue acts, the classification of post types and questions, the evaluation of the quality of posts, detection of topics, and user analysis. Searching in the forum can also improve the organization of data, for example identifying duplicate questions, or categorizing posts.

Discussion forums [39] are repositories of archived threads and response records that contain knowledge and experience in several topics. In addition, the responses to threads are very diverse, and depend on the learning style in a MOOC. Consequently, the answers which appear relevant can enrich the search for learners in a forum, and answer queries semantically.

In this article, we will create a ChatBot based on knowledge extracted from discussion forums in MOOCs and classified using artificial intelligence [22] according to the criteria of relevance.

Moreover, we present the knowledge-based ChatBot Framework and a novel method of extracting high-quality semantic thread-response pairs. The relevant thread-response

pairs make up the knowledge base of the ChatbBot Framework. The extraction of knowledge pairs is done by cascade processing and consists of several phases:

1) The preprocessing of raw data from learners' requests to identify their knowledge needs is carried out by NLP [21]
2) LDA: the allocation of latent dirichlet [7] will semantically classify thread responses according to topics and identify thread-response pairs

The classification is made by comparison with an ontology [38] which represents the knowledge domain in order to sort the pairs semantically. Rest of our article is as follows:

Section II relates to previous work, and we will spread their limits in relation to our work. In Section III we will present our Framework as well as its different components and their functionality. Section IV illustrates an experiment of our Framework. In the last section we will discuss the perspectives of our research.

2 Related Works

Most Automatic knowledge extraction through an unsupervised learning method is a new method for fully automatic ChatBot learning. Existing work for the automatic acquisition of ChatBots is mainly manual based on human annotation of datasets [22].

On the other hand, most ChatBots use templates, so in [23], we extracted knowledge from an online discussion forum, by automating the extraction of thread-response templates and building knowledge base for a Chatbot.

However, the extracted pairs are not linked by a structure like a graph or an ontology [38], which degrades the semantic quality of the ChatBot's knowledge. In particular, there is some work on the extraction of automatic knowledge.

In research [24], knowledge has been represented by an oriented graph, where the nodes represent the arguments and the arcs symbolize the conflicts between them. The conflicts arise when the chatbot takes a particular position in the dialog opposite to the user's position.

However, the corpus is not available which hinders the development of new chatbots, so in this article [24], we proposed a method to acquire a corpus of arguments in a graph structure using crowdsourcing.

In addition to checking for spelling errors, there was no other assessment of the quality of the arguments and no duplicate arguments were checked in the argument graph, which affected the quality of the results.

Another very interesting work [25] in automatic extraction is that of automatic extraction based on a model based on the combination of a rough set, and the theory of set learning for decision making. Responses are classified by classifiers based on rough set classifiers, and the results are drawn by vote on the output of the classifiers. Chatbot knowledge is mapped to associated responses.

Finally, the associated responses are selected as chatbot knowledge. An experiment was carried out on a childcare forum which gave satisfactory results for the method based on the rough set method.

This method possesses high recognition efficiency for related responses and the combination of learning together improves outcomes. Despite the good results from the application of rough sets and the theory of learning sets in data analysis and system modeling, the quality of the response is affected by several attributes.

In this work we considered only a few structural descriptors and a few content descriptors. In addition, the forums have different styles and formats which change the descriptors to extract, so we must consider the characteristics shared by all the forums to improve the portability of the algorithm. Another point to improve in this algorithm is the design of a storage structure, to improve the recovery efficiency of the dialog management module.

In addition, the work in [35] deals with the application of the LDA in a discussion forum, to discover the interest of the learners for the topical ones. Indeed, the experiment identified the post of serious and non-serious users and the discovery of user interest.

However, this work requires some improvements in terms of the multitude of topics in a response which makes the response fuzzy. Even so, this work is close to our discussion forum context, but does not deal with extraction of post replies and root replies according to the topics declared in a request.

3 LDA-Chatbot Framework

A forum is a tool for communication and discussion between people asynchronously online. Questions asked in the forums are subject to more than one answer or are factual questions such as a request for help. Users post messages related to the questions that matter to them to learn from community experiences.

Finding relevant information is difficult, and recommendation systems [31] [40] can help filling this gap by providing users with discussion threads related to their research interests.

A forum [25] [26] [27] is a tool for learner dialogue and online communication. A forum consists of several discussion sections with different topics discussed.

The user starts the discussion in a thread created at the beginning of the discussion (starting point) which is the starting post. Other users respond to the start-up post or comments from users already posted. As the user can ask questions by posting questions to the existing section.

In a section the threads are listed in chronological order. The term ‘thread’ in forums refers to threads in the discussion forum with all of their responses (and comments).

A thread contains a title, a start thread and a number of responses related to it. The thread title is the title of the root message posted by the starter thread to initiate the discussion. In the start thread, the user who starts a new discussion thread by writing the initial post. The user can view the list of responses in chronological order of a given thread, with information on the authors and the time of publication (time posting).

Thus, the structure of a discussion forum thread can be seen as a tree with an initial post at the top (start post), and responses to the post. Each post is placed under the post to which it responds.

The discussion forum has a large repository of thread archiving, and recorded responses, which contain great potential for knowledge on a given subject. In a context of learning by MOOC [30], there are various learners with different styles, which generate very varied responses in a forum, and convey diverse knowledge. Learners' posts reflect behavior in a MOOC [16]. The quality level of the thread's responses is an indicator of the level of knowledge buried in a response to a thread. The Knowledge in the threads can be of great value in the building a ChatBot in certain areas.

Given the number of subjects and the number of learners registered in a MOOC forum, which generate thread pages and thus knowledge of different levels, gives us a knowledge base for the creation of a semantic ChatBot.

Our ChatBot approach is a new approach to extract and present to the learner, the most relevant recommendations classified according to the degree of semantics carried by each response.

The extracted answers will enrich the ChatBot knowledge database. Indeed, our approach in this article is the development of a recommendation system [40] by ChatBot in a MOOC, based on the knowledge extracted in a discussion forum.

The conceptual framework of our ChatBot recommendation system is illustrated in Figure 1. The ChatBot semantic recommendation framework is divided into 5 phases:

1) First Phase: Receiving user-questions (Post a request in chatbot)
2) Second Phase: Analyzing the requests by ChatBot Production system (Preprocessing the request, extract keywords)
3) Third phase: Classification of the key requests by the LDA model (classification of the messages for each request's keyword)
4) Fourth phase : Mapping to domain ontology of MOOCs.(mapping the relevant threads containing the concepts with a domain ontology)
5) Fifth Phase: Recommending semantic items

The user asks a question to ChatBot which transmits it to the Processing Framework to find the relevant answers in the MOOC discussion forum. The recommendation process is started by making a request to the Discussion Forum database to send it all the {thread-response>} sets that contain the keywords of the request.

The posted messages can contain discussions between the learners relating to topics, subjects, or request for assistance, tutorials, articles, in fact any information on the teaching resources.

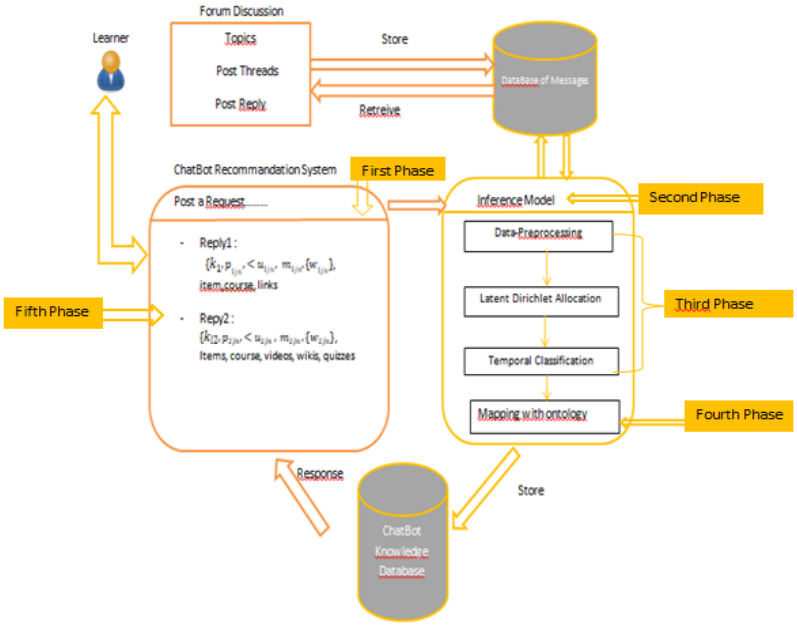

Fig. 1. The LDA-ChatBot Framework for Semantic Recommendation

On the other hand, all of the data pairs extracted from the {thread-response} database are in a raw format, which requires the intervention of the preprocessing process, to eliminate stop-words word-stemming [1], and all tags to prepare the data as an input.

Then, the preprocessed data goes into the LDA probabilistic model phase [7], to capture the latent messages in the posted messages and to model them by topics.

While the {thread-post} peers are classified by topics, we will list these message pairs by relevance, which results in the highest probability of belonging in the recommendation prediction phase. The messages will then be sorted according to the nearest neighbor's time metric which contains the messages sorted by relevance sent to the ChatBot module, which will store them in its knowledge database and then display them on the user interface. We will discuss the processes of each phase in the following sections.

3.1 Pretreatment phase

In this phase the data extracted from the forum database in the form {thread-response} are prepared for the modeling phase. During this phase we will eliminate the stop-words and stemming word.

Thus, we eliminate the stop words in each peer analyzed (parsed), which have no influence on the overall meaning of the message such as: he, the, it, is, to, wich, him, etc. as well as the special characters and the numbers are also eliminated. Stigmatization

carried out through converting each word of the message posted to its root (such as learning and learning will be brought back to their root 'learn') [28].

3.2 LDA for chatbot

The nature of LDA [7] is a probabilistic model used to discover a mixture of subjects and documents throughout the entire corpus. The model considers each document as a mixture of subjects which are hidden, and each subject is presented by a vocabulary of words which is observed in the document.

Note that there are two unknown parameters in the LDA model, the subject distributions, and the document distributions that must be estimated by the probabilistic model. The model also has latent variables which manage the assignment of words to subjects.

In LDA model, we will define a subject as similar words which are observed in a collection of textual documents. A subject model like LDA [7] receives a set of input documents, a predefined number of topics, and a priori fixed parameters. Then, LDA tends to define a set of topics to describe the entry documents. Discovered distributions are combined to be mapped to each document entry. The topics themselves can be viewed as a list of words classified from high to low relevance.

The basic goal of our research is modeling chatbot recommendation system by LDA. We consider the thread-response pairs as our corpus, and the topics are formed by the outputs of the preprocessing phase. We are looking through the LDA post for the distribution of the query words over the {thread, response} of the discussion forum.

Discussion forums have a structure composed of threads, replies, user, date and time of publication. Threads in the forum are classified by the name of the topics and the responses to these threads. We consider an online forum as a collection of threads $T=\{t_1, t_2, t_3, \dots\dots\dots, t_n\}$, each thread contains a root message posted by the starting thread r_0.

The response r_i is posted by a user u_i at a specific time m_i, and composed of a group of words$\{w_i\}$.

Thus, a thread T can be modeled by all the triples:

$$T = \{< u_0,\ m_0, \{w_0\} >, < u_1,\ m_1, \{w_1\} >, \dots\dots\dots\dots\dots\dots\dots < u_n,\ m_n, \{w_n\} > \},$$

With the first triplet is the basic response to the root message posted by the startup thread which is the title of the section. In our article, we consider the request sent by the learner to the chatbot as a list of keywords: $K=\{k_1, k_2, k_3, \dots\dots\dots, k_m\}$.

The keywords in this list form the set of K topics of the parametric probabilistic approach.

By applying the parametric LDA [7] probabilistic approach to the discussion forum corpus, the result will be a distribution of each keyword on the different forum threads.

Since the LDA model consists in inverting the defined generative process and learning the posterior distributions of the latent variables in the model taking into account the observed data, our system will model the topics in the thread corpus, as illustrated in Fig 2.

The notations of our model are defined as follows:

K : Number of topics retained from the learner's request

U : Number of users who posted in the forum

D: Number of threads posted
N_f: Number of post-roots in the forum
M_d: Number of responses posted in each Thread d
α,β : Dirichlet distribution hyperparameters
Φ_d : Distribution of topics for thread d
θ_k: Distribution of root messages posted for a topic k
w : Word of the root message posted
z : Assignment of post roots to the chosen topic
y: Response topic posted
{r}: Set of answers asked by the user
π_u: Topic's distributions for each user

Our model is composed of T topics, and Φ_d the distribution of topics for each thread d. The discussion forum is made up of sections and each section contains D thread.

Each thread d consists of root message posted composed of words w_d posted by the learner and which are the topics subjects of the request. Posted replies identified by M_d relate to replies to the posted root message.

Each topic k, has a distribution on the root messages posted θ_k, the distributions of topics for each user π_u , are symmetric distributions of dirichlet parameters α.

The algorithm of the LDA-Chatbot process is described as follows:

Draw the distribution of topics for each user: π_u = Dirichlet(α)

Draw the distribution of the topics for each thread d: Φ_d = Dirichlet (β)

We identify a section of threads $d = 1,2 \ldots\ldots D$ in the forum:

Draw the distribution of root messages posted for each topic $\theta_k = Dir(\alpha)$.

For each topic in the learner's request $k = 1,2 \ldots\ldots K$:

draw a $post - root_{k,n}$, : $Z_{k,n} \sim multinomial(\theta_k)$ with n=1,2…N the number of post-root in the forum and draw a topic $T_{k,n} = multinomial(\Phi_{z_{k,n}})$

draw a post-reply vector $\{r_{k,d}\} \sim multinomial(\pi_{u,k,d})$ and for each topic $j = 1,2 \ldots\ldots J_{d,k}$ draw post-reply $r_{k,d,j} \sim multinomial(\Phi_{r_{k,n}})$

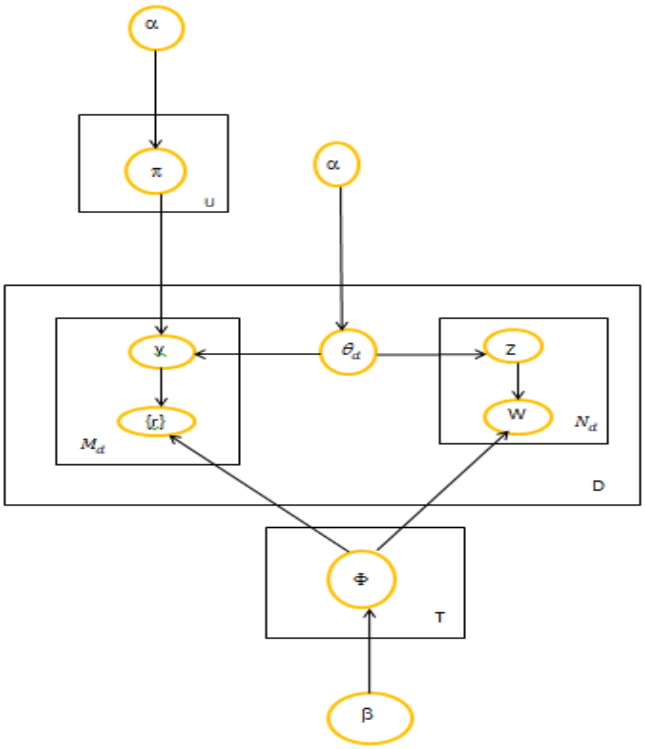

Fig. 2. The LDA-Chatbot Model

The inference in the LDA model is difficult to calculate, so we use Gibbs sampling [34] or the variation method, which are approximate algorithms to this problem. In this article we will apply Gibbs sampling to estimate the parameters. The sampling calculation (3) is divided into two distinct probabilities.

The probability of sampling which assigns a topic in the learner's request to a root post message is given by

$$\frac{N_d^k + C_d^k + \alpha}{N^d + C^d + K_\alpha} \tag{1}$$

and the probability which assigns a topic in the learner's request to a response message is given by

$$\frac{N_k^{pr_{d,n}} + \beta}{N_k + V_\beta} \tag{2}$$

$$P\left(z_{k,n} = r \middle| z_{\neg d,n}, k\right) \propto \frac{N_d^k + C_d^k + \alpha}{N^d + C^d + K_\alpha} \cdot \frac{N_k^{pr_{d,n}} + \beta}{N_k + V_\beta} \tag{3}$$

$z_{\neg d,n}$ denotes the topic assignment for all post-roots except $r_{d,n}$.

N_d^k is the number of thread root's post d assigned to topic k in ther user request, and N^d is the number of post-roots thread d in the forum assigned to topics.

C_d^k is the number of responses posted from thread d assigned to the topic k of the request, C^d d is the total number of responses posted from thread d.

$N_k^{pr_{d,n}}$ is the number of times when a post response is assigned to a topic k, and N_k is the number of times a post response is assigned to a topic k.

After performing Gibbs sampling, we can estimate the parameters of the latent variables by the following equations:

For the variable $\Phi : \Phi_k^v = \frac{N_k^d + \beta}{N_k + V_\beta}$; N_t^d is the number of time that a thread d is assigned to a topic k, et N_k is the total number of times that any thread is assigned to the topic k.

For the variable $\theta_d^k : \theta_d^k = \frac{N_d^k + C_d^k + \alpha}{N_d + C_d + K_\alpha}$

After convergence, the parameters of the new post can be calculated by the variables θ_d^k and Φ_k^v.

By applying the LDA model for a set of words in a request

$K = \{k_1, k_2, k_3, \dots\dots\dots, k_m\}$, we calculate the probabilities of distribution, that is to say the weights of each set for a keyword:

For k1 :

$$P_1 * < u_{110},\ m_{121}, \{w_{23}\} > + p_{21} * < u_{210},\ m_{141}, \{w_{265}\} > + \cdots \dots \dots p_n * < u_{110},\ m_{121}, \{w_{23}\} + \cdots$$

For k2:

$$P_1 * < u_{220},\ m_{240}, \{w_{533}\} > + p_{21} * < u_{210},\ m_{420}, \} > + \cdots \dots \dots p_l * < u_{10},\ m_1, \{w_{232}\} + \cdots$$

Until k_m keyword of the request:

$$P_1 * < u_{10},\ m_{21}, \{w_{36}\} > + p_{45} * < u_{98},\ m_{56}, \} > + \cdots \dots \dots p_j * < u_{25},\ m_{10}, \{w_{22}\} + \cdots$$

3.3 Ontology mapping

After having calculated the distribution of the threads on the keywords of the request which are the topics of the LDA model, we move on to the classification phase of the most relevant threads towards our request.

We search for each word in the list K, the associated threads in different distributions, then we extract the threads and classify them according to their weight according to the following algorithm:

$E_t = \emptyset$

For k_i *in* K :

$E_i = \emptyset$

For each k_j in K

Search in each distribution of k_j the tread wich contains k_i, $\{p_{ijn}, < u_{ijn},\ m_{ijn}, \{w_{ijn}\} >\}$

Fill the set $E_i = E_i \cup \{p_i, < u_i,\ m_i, \{w_i\} >\}$

End For
$E_t = E_t \cup \{k_i, p_{ijn}, < u_{ijn}, m_{ijn}, \{w_{ijn}\} >\}$
End For

After having filled the sets E_t for all k_i de K , which contains all the threads where word k_i is found and its probability distribution in the topics $K = \{k_1, k_2, k_3, \ldots\ldots\ldots, k_m\}$.

Now, we will classify these sets of threads by time constraint according to whether the closest time neighbors of the root message are the most relevant threads. So, we sort according to the time metric, and we get the closest neighbor to the root message which contains the topic k_i.

The sort is then sent to Chatbot for display to the learner, and since we are in the context of learning in a MOOC, we have enriched our Framework by an ontology which contains the concepts of educational resources.

Indeed, in Figure 3, we illustrate the extraction of the relevant threads containing the concepts, which will be mapped in ontology [29][39]. The instances of the concepts represent pedagogical materials [28].

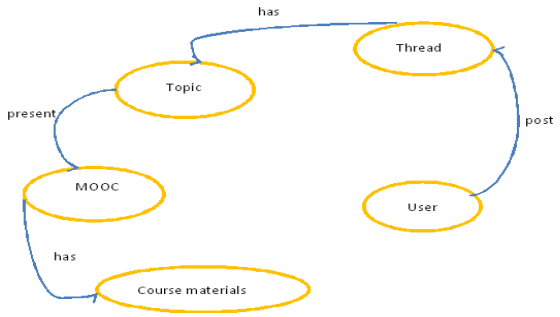

Fig. 3. Mapping of the concepts extracted with the ontology of the MOOC domain

4 Experimentation

In order to illustrate our approach to the development of semantic ChatBot, we will use the MOOC forum deployed in our Mohammadia School of Engineers. Messages and threads exchanged during learner interactions during the MOOC will be pre-processed initially. In the preprocessing phase we will eliminate stop words and stemmatize threads. Our goal is to extract the most relevant threads semantically according to the parameters of the learner's request in the chatbot.

For our model LDA, we put the Dirichlet priors α and β to be 0.1 and 0.01,1 respectively, since LDA with weak priors performs better in short texts [36]. With all values of α and β , Gibbs sampling is run for 5.000 iterations to obtain convergence.

So, the training and testing dataset contains 30% of the discussion forum threads, and responses to these threads range from a minimum of 5 to a maximum of 90.

The responses were labeled by three categories: 1) relevant response, 2) average relevant response, and 3) non-relevant response.

The automatic extraction by LDA, considers the keywords of the request, as being the topics of the latent allocation of dirichlet. We add a threshold of the first 10 threads as the most relevant, then the next 10 threads as medium and the last as irrelevant.

In addition, the extracted threads are the total number of 3000 messages, of which 1000 were used for training. During our experience, we used maximum likelihood to compare with our LDA-based approach as seen in Tab 1.

Table 1. Classification of the messages

Category	Phases	Number of messages
Category1: Relevant	LDA	1200
	Likelihood	300
Category 2: Medium relevance	LDA	800
	Likelihood	200
Category 3: Non relevant	LDA	300
	Likelihood	200

Indeed, we note that our approach is more consistent and has given very promising results. The category of relevant messages is the most significant with a total of 1200 messages with LDA and 300 with likelihood, the same for the category of average relevant and irrelevant. These measures confirm that our approach is very advantageous than the likelihood approach.

After having categorized the messages, we only retain those of the relevant category and we perform a temporal classification. The temporal classification favors the messages closest to the root message. In our experience we limited the messages to 10 first messages.

The timely sorted messages will then be mapped to domain ontology to extract the pedagogical resources which will be a support for the learner. A thread can contain various resources (learning activity, videos, documents, etc.).

For more comprehensive analysis, we evaluate our model using a similarity measurement metric [37] based on the semantics of the keywords sent by the learners to the model.

Similarity is measured by the distance between two concepts in ontology, that is to say the length of the path from the least common super term of the two terms to the root node [37].

This measurement takes into account the relation "of", to the nearest common ancestor to calculate the similarity. The similarity calculation formula between 2 concepts(c_1, c_2) takes the following form: $Sim(c_1, c_2)=\frac{2H}{D_1+D_2+2H}$

Note D_1et D_2 paths the closest c_1à c_2 and H the nearest road to c to root, respectively and c is the node with the least relationship is-a like ancestor node, which appeared to the lowest position in the ontology hierarchy.

By calculating the measures of similarity of the keywords of the learner's query in relation to the domain ontology, we then sort the different concepts according to the similarity between them.

In addition, the extracted threads are the total number of 3000 messages, of which 1000 were used for training. During our experience, we used maximum likelihood to compare with our LDA-based approach as seen in Tab 1.

We will only keep the first 5, 10, 15 concepts which have a high value of similarity, and we will apply our LDA-chatbot Framework model to those concepts. The number of extracted messages will remain invariable in the order of 3000 messages and we will only keep the relevant messages in our results:

Similarity	Phases	Number of messages
Category 1: 05	LDA	1000
	Likelihood	200
Category 2: 10	LDA	900
	Likelihood	700
Category 3: 15	LDA	150
	Likelihood	50

The results were interesting, as illustrated in Tab 2, the number of messages extracted by the LDA approach was higher than that extracted by likelihood as in the 1st experiment.

Moreover, the number of relevant extracted messages in category 1 is far ahead of category 2 and category 3, which means that the greater the similarity the more relevant the extracted messages.

However, the greater the number of concepts, the fewer relevant messages are extracted, which confirms the research results of [36] that LDA with weak priors performs better in short texts.

These results show the performance of our system in extracting relevant messages from a discussion forum and give very interesting results for queries with few keywords.

On other side, the greater the number of concepts, the fewer relevant messages are extracted, which confirms the research results of [36] that LDA with weak priors performs better in short texts, in fact our system is very adaptive for short texts than long text.

5 Conclusion

In this article we have presented a ChatBot Framework of semantic recommendation which present responses to the different requests of the learners sent in a natural language.

The answers extracted from the discussion forums are rich semantically, and translate the knowledge buried in the threads posted in the forum. The ChatBot in a MOOC environment is designed as a course facilitator and not a chat tool with learners.

This article discusses the value of semantic ChatBot for analyzing, exchanging, and sharing learners' knowledge and experiences. Our Framework is composed of 5 phases: Receiving user-questions, Analyzing the requests by ChatBot Production system, Classification of the key requests by the LDA model, Mapping to domain ontology of MOOC, Recommending semantic items.

The application of the LDA probabilistic model makes it possible to extract the relevant knowledge, according to a probabilistic model for each key word of the learner's request. The extracts threads will be drawn temporally based on timestamp of the message root.

Then, we enrich the responses, by mapping the concepts extracted with MOOC ontology. Experiments show that our model based on topic modeling (such as LDA) can be effective to build Production systems and semantic response in MOOC platforms.

In addition, the data treated in the phases of the Framework are massive data, so it consumes time and capacity. We must add a Big Data algorithm phase such as Map Reduce [31], SPARK ML [32] to our Framework in order to obtain better performance in time and energy consuming.

In the next works, we will exploit the combination of the LDA approach with the Hawkes process [34] for a better taking into account of the time factor, or the exploitation of the LSTM algorithm [33].

6 References

[1] M. Yan, P. Castro, P. Cheng, V. Ishakian, "Building a chatbot with serverless Computing," in Proc. the 1st International Workshop on Mashups of Things and APIs Article No. 5, 2016. https://doi.org/10.1145/3007203.3007217

[2] Y. Bi, K. Deng, and J. Cheng "A Keyword-Based Method for Measuring Sentence Similarity," in Proc. the 2017 ACM on Web Science Conference, pp 379-380, 2017 https://doi.org/10.1145/3091478.3098878

[3] G. D'Aniello, A. Gaeta, M Gaeta, S. Tomasiello "Self-regulated learning with approximate reasoning and situation awareness," Journal of Ambient Intelligence and Humanized Computing, pp. 1-14, 2016 https://doi.org/10.1007/s12652-016-0423-y

[4] Heffernan, N.T., Web-Based Evaluations Shsuicowing both Cognitive and Motivational Benefits of the Ms. Lindquist Tutor, in Artificial Intelligence in Education. 2003, IOS Press: Amsterdam. p. 115-122.

[5] Aleven, V., K. Koedinger, and K. Cross, Tutoring Answer Explanation Fosters Learn-ing with Understanding, in Artificial Intelligence in Education. 1999, IOS Press: Amsterdam. p. 199-206.

[6] Graesser, A.C., N.K. Person, and D. Harter, Teaching Tactics and Dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 2001. 12: p. 23-39.

[7] D. M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. JMLR.

[8] Chatterjee, P. and Nath, A., 2014. Massive Open Online Courses (MOOCs) in Higher Education – Unleashing the Potential in India. 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), pp. 256–260. https://doi.org/10.1109/mite.2014.7020283

[9] Bassi, R., Daradoumis, T., Xhafa, F., Caballe, S., and Sula, A., 2014. Software Agents in Large Scale Open E-learning: A Critical Component for the Future of Massive Online courses (MOOCs). 2014 International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 184–188. https://doi.org/10.1109/incos.2014.15

[10] Goh, O. S., Abd Ghani, M.K., Kumar, Y.J., Choo, Y.H., Muda, A.K., 2014. Massive Open Online Course (MOOC) with Learning Objects and Intelligent Agent Technologies. 2014 International Conference on IT Convergence and Security (ICITCS), pp. 1-4. https://doi.org/10.1109/icitcs.2014.7021786
[11] Laura Pappano. 2012. The Year of the MOOC. The New York Times 2, 12 (2012), 2012
[12] De Pietro, O. and G. Frontera, TutorBot: an application AIML based for Web-Learning. Advanced Technology for Learning, 2005. 2(1): p. 29-34. https://doi.org/10.2316/journal.208.2005.1.208-0835
[13] Samuel Coppey, Chatbot: Le pont entre clients et professions libérales, HES-SO Valais - Haute Ecole de Gestion, 2018.
[14] Mikic, F.A., Burguillo, J.C., Llamas, M., Rodríguez, D.A., & Rodríguez, E., 2009. CHARLIE: An AIML-based Chatterbot which Works as an Interface among INES and Humans. EAEEIE Annual Conference, pp. 1-6. https://doi.org/10.1109/eaeeie.2009.5335493
[15] Chatterbot tutorial. http://chatterbot.readthedocs.io. Dernière consultation: 2017-06-06.
[16] Goh, O. S., Ardil, C., Wong, W., Fung, C. C., 2007. Black-box Approach for Response Quality Evaluation of Conversational Agent Systems. International Journal of Computational Intelligence, 3 (3), pp. 195-203.
[17] Kauza, J., 2014. More Questions than Answers: Scratching at the Surface of MOOCs in Higher Education. In Invasion of the MOOCs: The Promises and Perils of Massive Open Online Courses (Krause, D. and Lowe, C.), pp. 105-113. South Carolina: Parlor Press.
[18] Samuels, B., 2014. Being Present in a University Writing Course: A Case Against MOOCs. In Invasion of the MOOCs: The Promises and Perils of Massive Open Online Courses (Krause, D. and Lowe, C.), pp. 68-72. South Carolina: Parlor Press.
[19] DA Kane, The role of chatbots in teaching and learning, - E-Learning and the Academic Library.
[20] De Pietro, O. and G. Frontera, TutorBot: an application AIML based for Web-Learning. Advanced Technology for Learning, 2005. 2(1): p. 29-34. https://doi.org/10.2316/journal.208.2005.1.208-0835
[21] Delphine Bernhard and Iryna Gurevych. Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA), pages 44–52. ACL, 2008. https://doi.org/10.3115/1631836.1631842
[22] B. A. Shawar and E. Atwell. Machine Learning from dialogue corpora to generate chatbots. In Expert Update journal, 6(3):25-29, 2003.
[23] J. Huang, M. Zhou, and D. Yang. Extracting chatbot knowledge from online discussion forums. In Proc. of the 20th International Joint Conference on Artificial Intelligence, pages 423–428, 2007.
[24] L. A. Chalaguine and A. Hunter. Knowledge acquisition and corpus for argumentation-based chatbots. In Proc. of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence, pages 1–14, 2019.
[25] Doris Hoogeveen, Li Wang, Timothy Baldwin and Karin M. Verspoor (2018), "Web Forum Retrieval and Text Analytics: A Survey", Foundations and Trends® in Information Retrieval: Vol. 12: No. 1, pp 1-163. https://doi.org/10.1561/1500000062
[26] Holtz, P., Kronberger, N., & Wagner, W. (2012). Analyzing internet forums: A practical guide. Journal of Media Psychology: Theories, Methods, and Applications, 24(2), 55–66. https://doi.org/10.1027/1864-1105/a000062
[27] Porter,M. F.(1980). An Algorithm for suffix stripping. Program, 14(3), 130-137.
[28] F. Colace, M. D. Santo, M. Lombardi, F. Pascale, and A. Pietrosanto, "Chatbot for E-Learning: A Case of Study," vol. 7, no. 5, p. 6, 2018.
[29] Y.T. Benjelloun,H.Abdeladim,N.EL FADDOULI, Machine Learning for Knowledge Construction in a MOOC Discussion Forum, International Journal of Innovative Technology and

Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020. https://doi.org/10.35940/ijitee.c8899.019320
[30] Holotescu, Carmen. "MOOCBuddy: a Chatbot for personalized learning with MOOCs." RoCHI (2016).
[31] Shim K. (2013) MapReduce Algorithms for Big Data Analysis. In: Madaan A., Kikuchi S., Bhalla S. (eds) Databases in Networked Information Systems. DNIS 2013. Lecture Notes in Computer Science, vol 7813. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37134-9_3
[32] Meng, Xiangrui et al. "MLlib: Machine Learning in Apache Spark." J. Mach. Learn. Res. 17 (2015): 34:1-34:7.
[33] Soutner D., Müller L. (2013) Application of LSTM Neural Networks in Language Modelling. In: Habernal I., Matoušek V. (eds) Text, Speech, and Dialogue. TSD 2013. Lecture Notes in Computer Science, vol 8082. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40585-3_14
[34] Darling, William M. 2011. "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 642–47.
[35] C. Chen and J. Ren, "Forum latent Dirichlet allocation for user interest discovery", Knowledge-Based Systems, vol. 126, pp. 1-7. (2017). https://doi.org/10.1016/j.knosys.2017.04.006
[36] Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu and Hui Xiong. Topic Modeling of Short Texts: A Pseudo-Document View, KDD, 2016. https://doi.org/10.1145/2939672.2939880
[37] Gan M, Dou X, Jiang R. From ontology to semantic similarity: calculation of ontology-based semantic similarity. Scientific World Journal. 2013; 2013:793091. https://doi.org/10.1155/2013/793091
[38] Hadioui, A., El Faddouli, N. E., Benjelloun Touimi, Y., & Mohammed, S. B. (2017). Machine learning based on big data extraction of massive educational knowledge. Interna tional Journal of Emerging Technologies in Learning, 12(11), 151–167. https://doi.org/10.3991/ijet.v12i11.7460
[39] Afify, M. K. (2019). The Influence of Group Size in the Asynchronous Online Discussions on the Development of Critical Thinking Skills, and on Improving Students' Performance in Online Discussion Forum. International Journal of Emerging Technologies in Learning (IJET), 14(05), 132–152. https://doi.org/10.3991/ijet.v14i05.9351
[40] M. El Mabrouk, S. Gaou, and M. Rtili, "Towards an intelligent hybrid recommendation system for e-learning platforms using data mining," International Journal of Emerging Technologies in Learning, vol. 12, no. 6, pp. 52–76, 2017. https://doi.org/10.3991/ijet.v12i06.6610

7 Authors

Yassine Benjelloun Touimi, Abdelladim Hadioui, Nour-eddine El Faddouli, and Samir Bennani are with RIME TEAM-Networking, Modeling and e-Learning- LRIE LaboratoryResearch in Computer Science and Education Laboratory at Mohammadia School Engineers (EMI) - Mohammed V University in Rabat Agdal AV. Ibn Sina Agdal Rabat BP. 765 Morocco.

Article submitted 2020-05-20. Resubmitted 2020-06-17. Final acceptance 2020-06-18. Final version published as submitted by the authors.