IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 26, 2020, accepted August 7, 2020, date of publication August 12, 2020, date of current version August 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016142

KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data

ADDI AIT-MLOUK AND LILI JIANG
Department of Computing Science, Umeå University, 901 87 Umeå, Sweden
Corresponding author: Lili Jiang (lili.jiang@umu.se)

This work was supported by Umeå University, Sweden, on federated database research.

ABSTRACT With the rapid progress of the semantic web, a huge amount of structured data has become available on the web in the form of knowledge bases (KBs). Making these data accessible and useful for end-users is one of the main objectives of chatbots over linked data. Building a chatbot over linked data raises different challenges, including user queries understanding, multiple knowledge base support, and multilingual aspect. To address these challenges, we first design and develop an architecture to provide an interactive user interface. Secondly, we propose a machine learning approach based on intent classification and natural language understanding to understand user intents and generate SPARQL queries. We especially process a new social network dataset (i.e., myPersonality) and add it to the existing knowledge bases to extend the chatbot capabilities by understanding analytical queries. The system can be extended with a new domain on-demand, flexible, multiple knowledge base, multilingual, and allows intuitive creation and execution of different tasks for an extensive range of topics. Furthermore, evaluation and application cases in the chatbot are provided to show how it facilitates interactive semantic data towards different real application scenarios and showcase the proposed approach for a knowledge graph and data-driven chatbot.

INDEX TERMS Linked data, Chatbot, SPARQL, intent classification, natural language understanding, myPersonality dataset.

I. INTRODUCTION

The use of chatbot is very popular since its inception in 1960. After two decades, research and development have seen impressive progress from Eliza 1960 to AI chatbots such as Siri 2010, Cortona, and google assistant. Early chatbot systems, such as Eliza [1], Parry [2], and Alice [3], were designed based on text conversation. A chatbot is a virtual agent able to assist users by providing instant responses to the instant question provided by the user. It is not just a conversational system; they can also carry out other tasks such as ordering, booking, customer care, and many other tasks. In the past several years, giant companies have invested in artificial intelligence and developed several chatbots, among them Apple's Siri,[1] Microsoft Cortana,[2] Google Assistant,[3] Facebook Messinger,[4] and Alexa.[5]

In the context of linked data, the main purpose of chatbot systems is to retrieve useful and relevant information from one or multiple knowledge bases (KBs) by using natural language understanding (NLU) and semantic web technologies. This purpose is generally addressed by transforming natural language into a SPARQL query. Many chatbot systems have been proposed, but they required a lot of training data, which is unavailable and expensive to create. Recently, with the growth development of linked data, increasing progress on chatbots have been seen in research and industry. However, they are still facing many challenges, including user queries

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

[1]https://www.apple.com/ios/siri/
[2]https://www.microsoft.com/en-us/cortana/
[3]https://assistant.google.com/
[4]https://developers. Facebook.com/blog/post/2016/04/12/
[5]https://developer.amazon.com/alexa/

understanding, intent classification, multilingual aspect, multiple knowledge base, and analytical queries understanding. In this paper, we propose a chatbot (KBot) that addresses some of the above challenges, and that can compete in terms of performances with existing linked data chatbots. Besides, we process one of the largest research databases in social science (i.e., the myPersonality corpus[6]), which was collected from over 6 million volunteers on Facebook (FB). The data was anonymized and sampled to share with registered scholars around the world. We use three of its data sources, including *demographic dataset, personality dataset, community detection*. Our key contributions are:

- We design and build KBot;
- We build a classifier for intent classification using a machine learning model (SVM);
- We provide analytical queries engine, including data exploration, which empowers users to explore myPersonality data via analytical queries (e.g., dominant political view, relationship status, and personality, personality traits, etc.);
- Insure the scalability, KBot is flexible by adding other knowledge bases, support new languages, and aiming at different tasks.

The remainder of this paper is organized as follows. Section II presents a survey of related works. Section III explains the architecture and describes the core techniques of the proposed chatbot system as well as analytical queries over myPersonality dataset. In Section IV, we demonstrate the results obtained and evaluation. Finally, section V concludes the work and outlines future work.

II. RELATED WORK

One of the early chatbots was ELIZA by Weizenbaum. It was implemented in the Massachusetts Institute of Technology to emulate a psychotherapist, it can communicate with humans based on hand-crafted scripts. This chatbot is inspired by the Turing test proposed in 1950 [4]. Later, researchers have proposed and developed multiple conversational systems for user interaction. These early chatbots are mainly based on text and pattern matching. Parry is another early chatbot developed by a psychiatrist Colby in 1972; it is still rule-based and implemented a crude model of the behavior of a person with paranoid schizophrenia based on concepts. Later, ALICE (Artificial Linguistic Internet Computer Entity) was introduced and developed by Wallace [3] using an artificial intelligence markup language (AIML)[7] to allow users to customize their chatbots. Rose [5] is another chatbot that uses a comprehensive natural language engine to recognize the meaning of the input sentence, and it combines chat script with question understanding. The other type of intelligent conversational system is smart personal assistant, among theme Microsoft Cortona in 2003, which can control windows mobile by voice, Siri, released by Apple in 2012 to control devices based on macOS, iOS, watchOS, and tvOS by voice. Google's assistant, released by Google in 2016 for mobile and smart home devices. Alexa released by Amazon as a virtual assistant.

Generally, a chatbot can be categorized into three categories, educational, healthcare, and business. They are based on the knowledge bases from those domains to provide consistent support for the user. Educational chatbot help students answer specific education-related queries such as MedChatbot [6] for medical students, based on the open-source AIML. In this chatbot, authors deploy a widely available Unified Medical Language System (UMLS) as the domain knowledge source for generating responses and translating natural language queries into relevant SQL queries. These SQL queries are run against the knowledge base, and results returned to the user in the natural dialogue. CSIEC [7] Computer Simulation in Educational Communication, a system with newly developed multiple functions for English instruction that can chat interactively in English with the English learners. It generates response according to the user input, the dialogue context, common sense knowledge, and inference knowledge. Freudbot [8] for the psychology domain, it was designed to chat in the first person about his theories, concepts, and biographical events using a resource that contained dictionary-type definitions of Freudian terms and concepts. Other chatbots for FAQs for university-related questions are proposed among them [9]. Healthcare chatbots support patients with an answer to specific questions related to healthcare such as Mamabot for supporting women and families during pregnancy [10] exploiting AI-based chatbot systems to understand and respond to patient needs. Also, Pharmabot [11] in 2015, is a pediatric generic medicine consultant chatbot designed to prescribe, suggest, and give information on generic medicines for children. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation [12] that suggests a conversational service for psychiatric counseling. This chatbot adapted methodologies to understand counseling contents based on high-level natural language understanding (NLU), and emotion recognition based on a multi-modal approach. Divya *et al.* [13] proposed a medical chatbot. The idea is to create a text-to-text chatbot that engages patients in conversation about their medical issues and provides basic information and diagnosis based on their symptoms. Hence, people will have an idea about their health and have the right protection. To improve the quality of life for young adults with food allergies, AllergyBot [14] was proposed as an intelligent and humane chatbot that provides restaurants' allergy accommodation information based on users' allergens. Besides, chatbots from the business domain are also proposed to support customers and improve companies' services. In e-commerce systems, SuperAgent [15] is a customer service chatbot that leverages large-scale and publicly available e-commerce data. It takes advantage of data from in-page product descriptions as well as user-generated content from e-commerce websites, which

[6]http://mypersonality.org
[7]http://www.aiml.foundation/

is more practical and cost-effective when answering repetitive questions. Moreover, IBM proposed a conversational system to generate responses for users' requests on social media automatically. It is integrated with state-of-the-art deep learning techniques and trained by nearly 1M Twitter conversations between users and agents from over 60 brands [16]. This conversational system adopts a word embedding method, word2vec neural network language model [17], to learn distributed representations of words from the customer.

In the context of the semantic web, chatbots systems use structured and linked data to support conversation in different tasks (Question answering systems, FAQ, etc.). Among them, OnBot [18], which is a new ontology-based approach proposed to model and operate chatbots. OnBot uses appropriate mapping techniques to transform ontologies and knowledge into a relational database to drive its chats. Bota [19] chatbot explores the challenges of creating a conversational agent that aims to simulate friendly conversations using the Egyptian Arabic dialect. Besides, AliMeChat [9], is an open-domain chatbot engine that integrates the joint results of information retrieval (IR) and sequence to sequence (Seq2Seq) based generation models. Octopus [20] a multi-agent-based on eight sub-multi-agent systems, namely core system, GUI system, natural language processing system, communication system, learning system, action system, searching system, and data access system to handle its capabilities. SOGO [21] is a semi-automatic social, intelligent negotiation dialogue system that interweaves task utterance with conversational strategies to engage human users in negotiation. La Liga [22], is another social chatbot for the football domain, aims at answering a wide variety of questions related to the Spanish football league. La Liga is deployed as a Slack client for text-based input interaction with users and used NLU block that trained to extract the intents and associated entities related to user's questions about football players, teams, and trainers. The information for the entities is obtained by making SPARQL queries to the Wikidata knowledge base site in real-time.

Many open-source frameworks that assist in creating conversational engines are proposed, among theme Microsoft Bot, Facebook Messenger, Google Assistant, and Amazon Lex. These frameworks build and connect intelligent conversation engines to interact with customers naturally wherever they are by taking advantage of the wide range of users (Facebook: 2.6 Billion). Besides, they are highly customizable in terms of real scenarios with third-party data. However, most of the systems are limited in terms of design, privacy, ethical issues, and natural question understanding. Many issues are addressed, while some others are not addressed at all. This paper aims to present our proposed chatbot over linked data that takes advantage of large-scale, publicly available knowledge bases (DBpedia, Wikidata, myPersonality) on one side. On the other side, it can address user intent classification, multilingual aspect, and handle analytical user queries using a newly established knowledge base (myPersonality).

III. OVERVIEW OF KBot

Our proposed approach is based on a modular approach, as shown in Figure 1, and takes advantage of semantic web techniques, knowledge graph, and machine learning. The proposed chatbot can handle different tasks (e.g., analytical queries, FAQs, etc.), gathering information from multiple sources (KBs, web services), and presenting them in the form of knowledge card (info-boxes that shows users' responses and displays only important attributes about user query). To enhance usability, users can interact with the system by using a chit-chat system or voice-based messages. The proposed chatbot is developed using Flask framework[8] and can run on standalone or distributed mode to improve response time of information retrieval. The queries and user feedback are stored in a database in an anonymous way for continuous learning and future improvement. More details about the working mechanism of the proposed chatbot will be explained in the following subsections.

Algorithm 1 Pseudo-Code of Response Retrieval Algorithm

```
Input  : Question, Endpoint
Output: Response
Entities ← ∅
Keywords ← ∅
Input ← Question
Query ← Process_Input(Input)
Intent ← Get_Intent(Query)
foreach Word W_i ∈ Question do
    Entity ← Ner_Tagger(W_i)
    Entities ← Entities + Entity
end
if Entities equal 0 then
    Keywords ← Parser(Query)
    Response ← Get_Response(Keywords)
else
    Query ← QuestionToSparql(Entities, Intent)
    Response ← Get_Response(Endpoint, Query)
end
return Response
```

A. NATURAL LANGUAGE UNDERSTANDING (NLU)

The interaction begins with a text/speech user query, which will be processed by the *'NLU'* module. Speech Known as automatic speech recognition (ASR) is one of the core techniques of computational linguistics that develops methodologies that enable the recognition of spoken language into text by computers. This technique can help disabled people who cannot use other devices to interact with chatbots, and it also improves typing speed. *'NLU'* module firstly detects the language of the user query. Afterward, it classifies user

[8]https://flask.palletsprojects.com

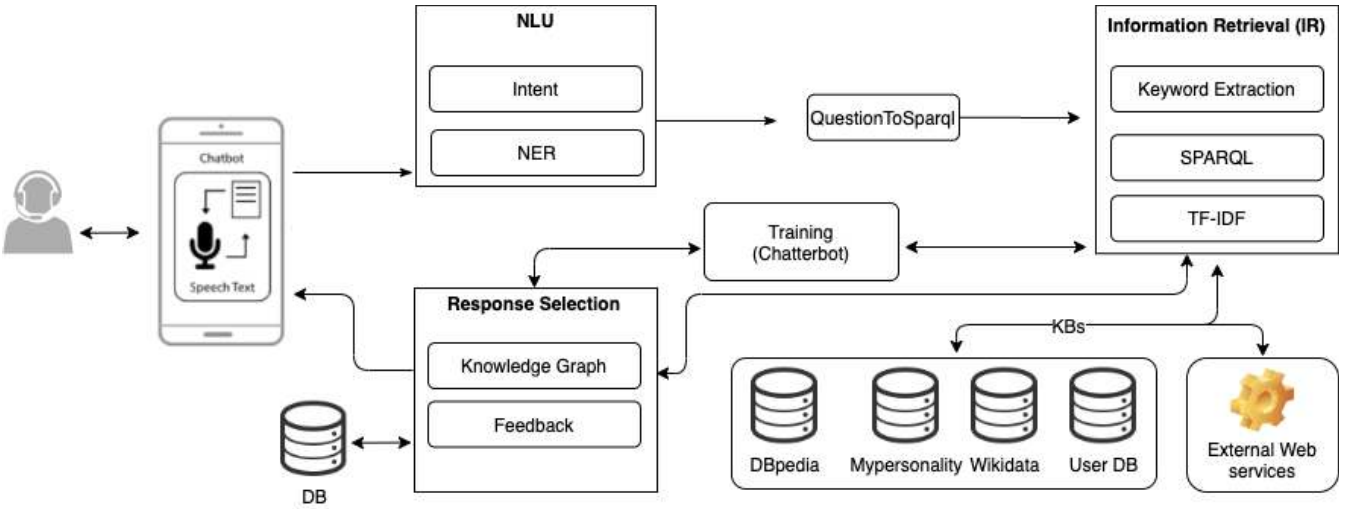

FIGURE 1. Architecture overview of KBot.

intents using Support Vector Machine (SVM) and parses the user query's structure by using a specific parser in combination with a regular expression (Regex). After processing the user query, the *'NER'* sub-module is used to extract the named entities mentioned in the query. This sub-module is considered as the key to understanding natural language questions. If the user query belongs to factoid questions class, the sub-module *'QuestionTosparql'* converts the user question to SPARQL query and retrieves the relevant response from KBs (i.e., DBpedia, Wikidata) and third-party services (OpenStreetMap, API, etc.). If the query belongs to the social network dataset, *'QuestionTosparql'* converts the question to the SPARQL query and retrieves a response from myPersonality dataset. Users can query myPersonality dataset structure, attributes, basic statistics, and analytical queries (e.g., dominant political view, relationship status, and personality, personality traits, etc.). Once, the response is retrieved, the *'Response Selection'* module selects a relevant answer and presents it to the user in a rich knowledge panel.

1) MULTILINGUAL

Most existing chatbots and conversational systems have been developed for English users, given a large number of English resources available on the web. There is always a need to handle multilingual queries on conversational systems and chatbots to serve a wide range of users. In our proposed KBot, we used a langdetect library [9] to detect the language from the user question automatically. This module can analyze user questions and detect the language automatically. Afterward, the detected language (e.g., Swedish, Spanish, Arabic, etc.) will be used by *'Response Retrieval'* to retrieve the relevant answer; otherwise, the answer will be in English since the most covered content in KBs is in English. For instance, if a user asks a question in French "C'est qui Albert Einstein?", *'NLU'* module will detect that the user query is posted in French, and *'QuestionToSparql'* will retrieve and filter only the answer in French.

[9]https://pypi.org/project/langdetect/

2) INTENT CLASSIFICATION

Intent classification is usually the first stage in conversational systems. It is the process of mapping queries to a predefined class, and it aims to facilitate user query understanding. Query classes can raise different extraction strategies. The strategy used to search the answer for a query like "Who is Alan Turin" is probably different from the strategy used to search for a question like "Who invented the Turing machine". The first query is about description more specifically about the definition of "Alan Turing", while the second one expects the name of a person, which is "Alan Turing". To address this critical issue of distinguishing query types, we used a Support Vector Machine (SVM) [23] based query classification aiming to be language and domain-independent. We specifically focused on the user query classification part, and the purpose is to classify a given user query into predefined categories. This classification will help to identify user intention before generating SPARQL queries. In this context, we use a supervised machine learning model called support vector classifier (SVC) that uses classification algorithms for two-group classification. SVC takes training data points and outputs the hyperplane (decision boundary) that best separates two classes of samples. SVC has demonstrated to perform well with high dimensional data used in many NLP tasks, including text classification.

Given a training vectors $x_i \in R^p$, $i = \{1 \dots n\}$, in two classes, and a vector $y \in \{1, -1\}$, SVC solves the following problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + c \sum_{i=1}^{n} \zeta_i$$
$$\text{Subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i$$
$$\zeta_i \geq 0, \quad i = 1, \dots, n$$

Its dual is :

$$\min_a \frac{1}{2} a^T Q a - e^T a$$

Subject to $y_T a = 0, \quad 0 \leq a_i \leq C, \ i = 1, \ldots, n$

where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semi-definite matrix, $Q_{ij} = y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here, training vectors are implicitly mapped into a higher dimensional space by the function ϕ, (See more details in evaluation section).

3) NAMED ENTITY RECOGNITION (NER)

NER is one of the main components of the proposed KBot, which is employed to extract information about different types of entities (PERSON, GPE, ORG, etc.), relationships, or events by using tokenization and part-of-speech tagging (POS) approaches. For a given user query, we extract entities based on user intent then generate SPARQL queries to facilitate the exploitation of linked data (e.g., DBpedia, Wikidata, myPersonality, etc.) and use their strength to retrieve relevant answers. For instance, let $Q =$ *"Who is Alan Turing?"* denote a user query; the KBot will identify the class of question as "PERSON" and extract *"Alan Turing"* as the main entity. Afterwards, it generates the SPARQL query that gathers information about this entity from knowledge bases and presents it in a knowledge panel. Figure 2 presents a knowledge graph for the extracted entity from the user query.

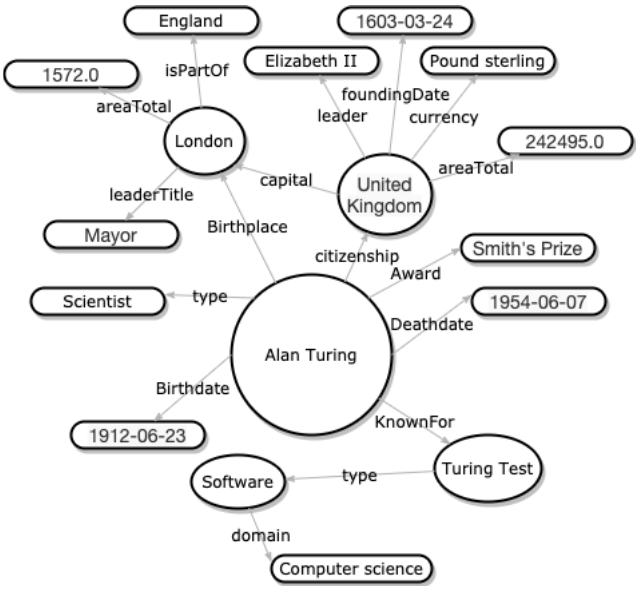

FIGURE 2. Knowledge graph example.

B. INFORMATION RETRIEVALS

In this sub-section, we present the various techniques used to retrieve the answer for a given query, including keyword extraction, Query generation, multiple knowledge bases, analytical queries, and text summarization (TF-IDF).

1) KEYWORD EXTRACTION (KE)

Keyword extraction is a text analysis technique with the automatic identification of terms that best describe a text document's subject. It helps summarize the content of a text and recognize the main topics which are being discussed. In this step, we use a specific parser with stop words and a regular expression to allow the chatbot to lift the most important words from user queries.

2) QUERY GENERATION (SPARQL)

In this step, we construct a set of SPARQL queries after processing and understanding the user query. These SPARQL queries represent a possible interpretation of user queries within the given KBs (DBpedia, Wikidata, myPersonality). The main objective is to generate a possible queries containing information about user queries. An example of a generated query is given in Figure 3, and the main challenge is to construct a SPARQL query from user question efficiently and query multiple knowledge bases according to user intent to retrieve a result-set.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT  ?comment ?birthdate ?birthplace ?thumbnail ?name
WHERE {
  ?x0 rdf:type foaf:Person.
  ?x0 rdfs:label "Alan Turing"@en.
  ?x0 rdfs:comment ?comment.
  FILTER (lang(?comment) = "en").
  ?x0 dbpedia:birthDate ?birthdate.
  ?x0 dbpedia:birthPlace ?birthplace.
  ?x0 dbo:thumbnail ?thumbnail.
  ?x0 foaf:name ?name.
} LIMIT 1
```

FIGURE 3. Example of generated SPARQL query for user question "Hello, give me information about Alan Turing.

3) MULTIPLE KNOWLEDGE BASE (KBs)

Many approaches use training data or focus on a unique and specific KB (e.g., DBpedia, Wikidata, etc.) to retrieve the response. These approaches are limited; they can retrieve responses from those KB without giving the user a chance to consider other KBs and can be limited in terms of using language and answers efficiency. In this paper, besides existing KBs, we processed an additional KB called myPersonality (Facebook dataset) consisting of social networks data as an extended and complemented knowledge based. KBot explores and analyzed it to find answers to some questions that can not be found in other knowledge bases, especially for psychological and social science researchers. It is one of the largest social science research databases, collected from over 6 million volunteers on Facebook (FB). myPersonality (Figure 4) was a Facebook App that allowed its users to participate in psychological research. It was created by David Stillwell in 2007 to share a personality questionnaire with Facebook users. The data was prepared and federated [24] to a virtual database (VDB), indexed in Apache Jena server

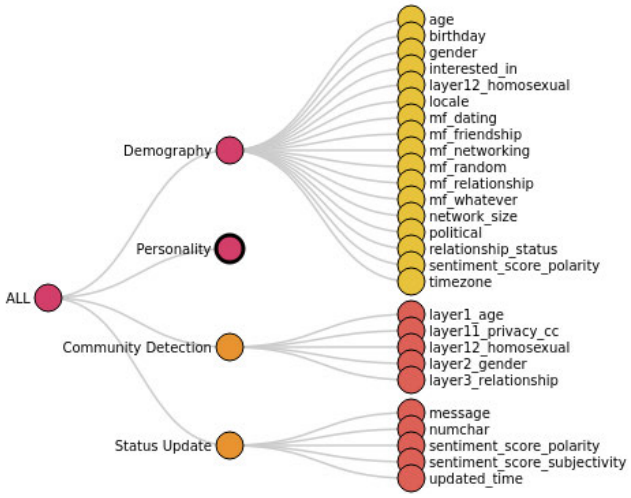

FIGURE 4. myPersonality dataset [24].

and accessible via an Endpoint to support semantic web and linked data applications. This VDB is anonymized and sampled to be shared with registered scholars around the world. In this paper, we used three of its data sources, including *demographic dataset, personality dataset, and community detection*. Moreover, we used action based on rules to handle user analytical queries, the pseudo-code is given in the Algorithm 2.

Algorithm 2 Action Based-Rules on myPersonality

Input : Query, Endpoint
Output: Response

Input ← *Query*
Query ← *Process_Input*(*Input*)
Intent ← *Get_Intent*(*Query*)
if *Intent in myPersonality* **then**
  keyword ← *Keyword_Matching*(*Intent*)
  Response ← *Data_Structure*(*keyword*, *myPersonality*)
  return Response
else
  Response ← *Get_Random_Answer*()
  return Response
end

4) ANALYTICAL QUERIES ON myPersonality DATASET

A large number of resources available on myPersonality dataset [25] and elsewhere may require further processing and analysis to generate valid answers through data analysis. The goal is to demonstrate modeling and efficient analysis to infer trends and establish correlations according to the user intents. Some of the analytical queries that are meaningful for myPersonality are shown below and presented in Figure 7.

Political Views:

1) How the user political view (e.g., democrat, does not care, republican) varies in the dataset?

Relationship Status:

1) How does relationship status (e.g., single, in a relationship, and married) vary?
2) How does the relationship status affect the personality traits of an individual?

Personality Analysis:

1) How does the personality traits evolve with age?

5) TEXT SUMMARIZATION TF-IDF

With the huge amount of semantic data in knowledge bases, which is mostly textual data, there is a need to develop automatic text summarization techniques that easily allow users to get insights. Most of the information is redundant, insignificant, and may not convey the intended meaning. For instance, if the user searches for specific information about entities (Google, Alan Turing, etc.) in linked data abstract and page description (DBpedia, Wikidata, etc.), he may have to dig through its content and spend unnecessary time before getting the intended information. In this context, we used NLTK[10] library to summarize and improve the result using TF-IDF; this can extract useful and relevant information. Text summarization using TF-IDF refers to the process of shortening the long text to only the main points outlined in the document. TF-IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). TF is the number of times a word appears in a document divided by the number of words in the document, while IDF is the log of the number of documents divided by the number of documents containing the word w. An example is given in Table 1.

TABLE 1. Example of answer to the user query using TF-IDF.

Query	Where is Sweden located geographically?
Answer	Approximately 85 of the population lives in urban areas. Sweden is part of the geographical area of Fennoscandia. The capital city is Stockholm, which is also the most populous city in the country. Legislative power is vested in the 349-member unicameral Riksdag. Executive power is exercised by the government chaired by the prime minister.

C. RESPONSE SELECTION (RS)

Once the response is retrieved, the Response Selection module gathers all retrieved information about the user query and presents it to the user in the knowledge graph.

1) KNOWLEDGE GRAPH

The knowledge graph represents a collection of interlinked entities, real-world objects, and events. It allows both people and computers to process them efficiently. In the proposed chatbot, the knowledge graph is typically built on top of the existing modules to link all retrieved information, combining structured or unstructured information, and presenting it as a valid knowledge panel to the user in a clear and structured

[10]https://www.nltk.org/

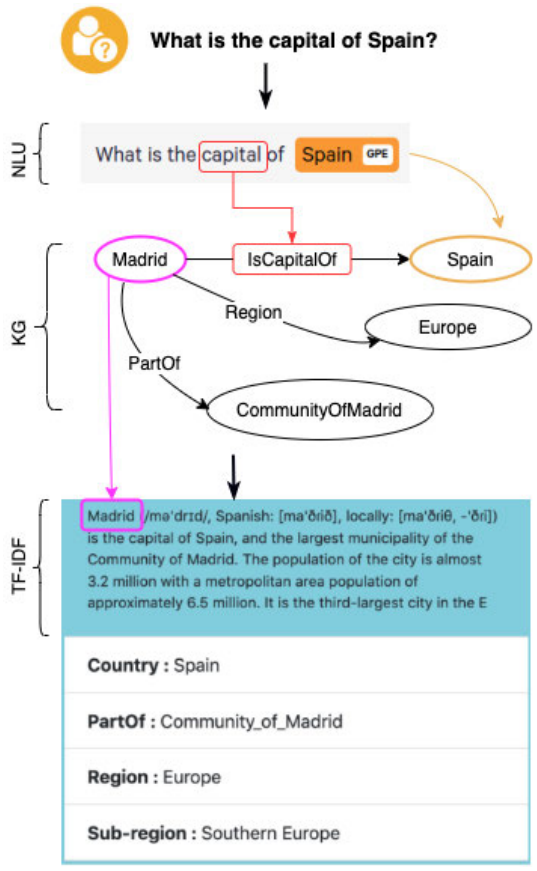

FIGURE 5. An example of response based on the knowledge graph.

way. For instance, let $Q =$*"What is the capital of Spain"* denote a user query, *"query understanding"* module will process the query and extract the entity *"Spain"* as a GPE and extract *"capital"* as a predicate, the generated query will take the tuple "<Spain, capital>" and return the triplet "<Spain, Capital, Madrid>" with other relevant information (e.g., Part of, Region, Sub-region), for more details see Figure 5. Moreover, the KBot uses geographical data to display the map (OpenStreetMap) of GPE entities; this will enrich its capabilities to answer questions that may contain information about geographical entities.

2) FEEDBACK AND CONTINUOUS LEARNING

A database layer is used to store user queries and feedback, which will be collectively analyzed to enhance and improve the performances for further usage. Besides, to extend the KBot to a wide range of topics, we used a training module *Chatterbot* to train the chatbot on specific tasks, such as English conversation, social networks conversation, FAQs, and mathematical queries. This module is based on an existing chatbot framework called Chatterbot.[11] Many other modules and libraries can be easily used closely with KBot to improve its capabilities and cover other tasks.

[11] https://chatterbot.readthedocs.io/

D. IMPLEMENTATION AND DEMO ENVIRONMENT

Designing and developing a chatbot is not a trivial task; it requires various design techniques and interactive chit-chat. Third-party frameworks have been proposed for a chatbot; it refers to open source building blocks that help developers community to build their conversation engines, among them Microsoft Bot, Facebook Messenger, Google Assistant, and Amazon Lex. These frameworks have some limitations in designing interfaces and advanced research techniques. Starting from this, we chose to design and develop our proposed KBot system over linked data using machine learning, NLP, QA, and knowledge graph. We developed interactive and user-friendly interfaces using the Flask framework. The proposed system has the following features:

- Robust: robust enough to deal with natural language questions and relevant keywords;
- Multilingual: supports multiple languages, including English and French, and can be adapted to a new language easily;
- Standalone: multiple platform (i.e., guarantee for low disk and memory footprint). It can be run on a standard laptop having two cores and 2GB of RAM;
- Real-time: can answer a user question in real-time;
- Precision and recall: achieve competitive performance compared with other linked data chatbots (see experiment and evaluation section).

The proposed system is accessible from different platforms to engage a wide range of users, and it is also optimized for both desktop and mobile. Figure 6 present some examples for querying linked data through two user queries *("Where is Sweden located geographically?", and "What is the capital of Spain?")*. Besides, Figure 7 present other examples about analytical queries from myPersonality dataset *("How does the relationship status vary?", "How does the political view vary?")*. For more information, a live demo can be found at (https://youtu.be/vcaNyGBNjbI).

IV. EVALUATION

A. INTENT CLASSIFICATION TASK

To correctly interact with the user in chatbot systems, we need to understand what the user asks for. Intent classification, (i.e., classifying the query into several categories) can suggest plausible constraints to get an answer. For example, if the chatbot understands that the query "Hello, Who is Alan Turing" is about a person's name, the confusion will be significantly reduced, and the answer will be relevant. The accuracy of the intent classification is crucial to the overall performance of the chatbot. To train the model using SVM, we follow the question taxonomy proposed in [26], which contains six coarse categories (i.e., entities, description, abbreviation, human, numerical, location) and 50 fine-grained question categories, as shown in Table 2. Usually, the number of question categories is less than 20. However, a fine-grained category definition is more beneficial in verifying the relevant answers. To simplify the experiments,

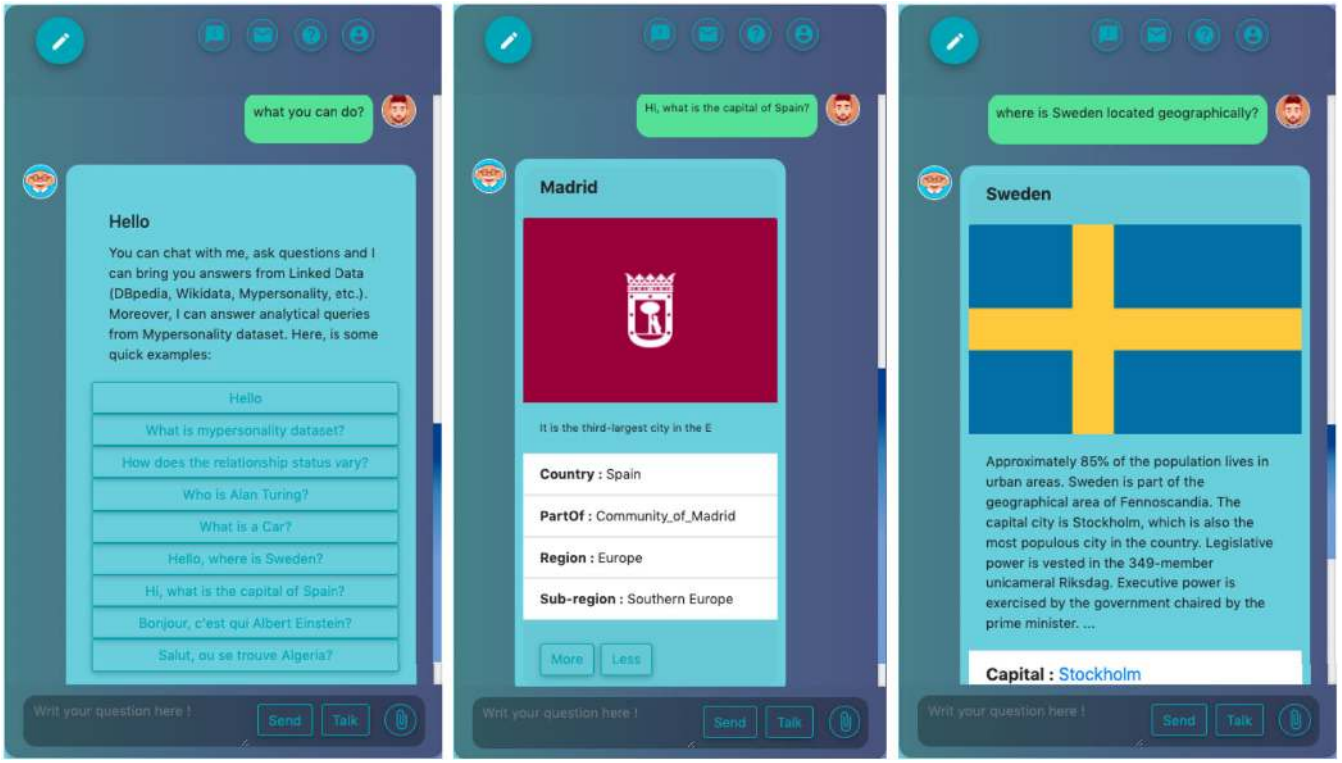

FIGURE 6. Use cases for natural language queries over linked data.

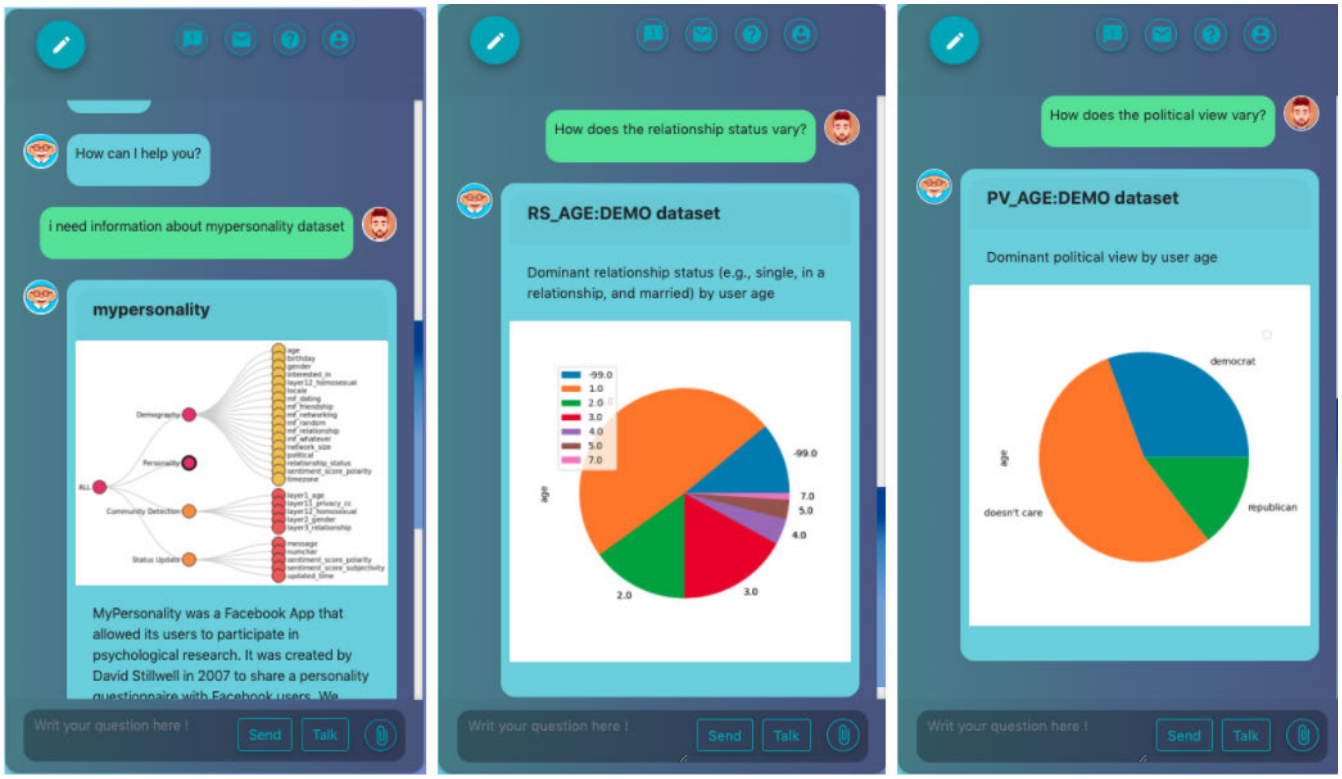

FIGURE 7. Use cases for analytical queries.

we assume that one question belongs only to one category, and the query is labeled with its most probable class.

TABLE 2. The coarse and fine-grained question categories.

Category	Fine category
ENTY	animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word.
DESC	definition, description, manner, reason
ABBR	abbreviation, expansion.
HUM	description, group, individual, title.
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight.
LOC	city, country, mountain, other, state.

UIUC has manually labeled the publicly available training dataset used in this paper according to the coarse and fine-grained categories. There are about 5500 labeled questions. Our testing dataset contains 100 queries and manually labeled according to our question hierarchy. We used Precision, Recall, and F-measure metrics for evaluation, the predicted intent was compared the labels, truth, and the evaluation result is given in Table 3. This table presents the classifier performance (precision, recall, and f-measure) using SVM; the result shows that SVM correctly identified the question classes. More specifically, SVM had 100% recall for class type LOC and similar recall for classes like NUM, HUM, and DESC. Weighted avg compute F1 for each label, and returns the average considering the proportion for each label in the dataset. The intent classification task is sensitive to the training data related to some classes and the language model[12] used during the training.

TABLE 3. Evaluation metrics for coarse classes.

Classes	Precision	Recall	F1-score	Support
ABBR	1.00	0.33	0.50	3
DESC	0.54	0.95	0.69	21
ENTY	0.12	0.12	0.12	8
HUM	0.88	0.93	0.90	30
LOC	1.00	0.35	0.51	26
NUM	0.85	0.92	0.88	12
W eighted avg	0.78	0.70	0.68	100

B. PERFORMANCE AND EFFECTIVENESS OF KBot

As shown in Table 4, we present the characteristics of KBot and some of the online available state-of-the-art chatbots (DBpedia chatbot, Open Data Chatbot, Open Data Assistant), all these chatbots makes use of linked data. The first column specifies the chatbot name together with the language it processed in case of multilingual queries, Knowledge Base states for the knowledge base used to provide an answer.

To evaluate the performance and effectiveness of KBot in terms of response to user queries, we used precision, recall, and F-measure. In this case, the precision is the ratio of the number of pertinent answers found over the total number of answers found. While the recall is the ratio of the number of pertinent answers found over the total number of relevant answers.

[12]https://spacy.io/usage/models

TABLE 4. Characteristics of the online available chatbots.

Chatbot	Language	Knowledge Base
DBpedia [28]	En	DBpedia
ODC [29]	En	DBpedia
ODA [30]	En	DBpedia
KBot	En/ Fr	DBpedia, Wikidata, myPersonality

TABLE 5. Results for each of the participating systems.

Chatbot	Total	Right	Wrong	P	R	F-measure
DBpedia	56	4	6	0.4	0.06	0.1
ODC	56	3	2	0.6	0.05	0.09
ODA	56	2	3	0.4	0.03	0.05
KBot	**56**	**19**	**3**	**0.8**	**0.3**	**0.44**

Table 5 reports the results obtained by the participating systems on the multilingual and hybrid queries, respectively. The first column specifies the chatbot name, Total states for the number of the queries; Right specifies the number of these questions answered correctly, Wrong states for the number of the queries the system provides the wrong answer. Recall, Precision, and F-measure report the metrics concerning the evaluation of the overall system. For example, with a total of fifty-six queries, KBot provides nineteen relevant answers, while other state-of-the-art provide less than five relevant answers. Meanwhile, KBot had (0.3) recall and outperformed other systems. These results indicate that in terms of precision, recall, and F-measure, KBot had better performance. The results also validate that combining NLP with linked data, especially myPersonality for analytical queries, improved the performances and enabled the machine learning algorithms (SVM) to understand natural language queries better and retrieve relevant answers.

The evaluation results show that our proposed approach achieves much better overall performance and outperforms the state-of-the-art systems over linked data. Based on that, the integration of NLP and intent classification within the queries understanding process over linked data performs well and produces a relevant response to user queries. After eliminating the non-interesting tokens from user query using a specific parser. 'QuestionTosparql' module generates SPARQL queries to retrieve relevant information from available knowledge bases (DBpedia, Wikidata, myPersonality, etc.).

C. DISCUSSION

In literature, one of the main issues of linked data is insufficiently integrated chatbot and similar systems built on top of semantic web technologies; most solutions are focused on a rule and AI-based chatbots. However, it ignored personalized knowledge bases and the strength of linked data and the semantic web technologies. To overcome some of these issues, we proposed KBot, a comprehensive open-access chatbot by exploiting the potential of semantic web technologies(RDF, SPARQL, etc.), federated database, and natural language understanding. For instance, it supports many tasks, including dialogue management, intent classification, FAQs,

question answering, analytical queries, and data exploration. With the integration of the myPersonality, KBot can answer some research queries and address particular use cases such as personality and sentiment analysis.

In summary, KBot contributes to a better understanding of user queries in the context of linked data (DBpedia, myPersonality dataset) by answering different user queries. The proposed KBot has the following three significant strengths: 1) the overall system provides an interactive user interface for dialogue management that facilitates user interaction; 2) Multiple knowledge base and multilingual with the integration of new social knowledge base (myPersobality) that support analytical queries; 3) the overall KBot provide open-access to involve a wide range of users and help researchers from social science to interact with myPersonality dataset using analytical queries.

V. CONCLUSION

In this paper, we proposed a knowledge graph-based chatbot system over linked data, optimized for community interaction. The proposed KBot system takes advantage of large-scale, publicly available knowledge bases, multilingual, speech to text, and external APIs. Besides, KBot leverages the technologies of machine learning and natural language understanding, including named entity recognition, factoid, and recurrent questions, as well as dialogue management. Usability analysis shows that the proposed KBot has improved the end-to-end user experience in terms of interactive question answering and performance. It is more convenient for information retrieval, information acquisition, intent classification, query understanding, and continuous learning.

The future work will be adding more text-based data sources with privacy preservation, answers generation-based, extending it to more knowledge bases and other languages, and integrating with third-party services (Slack, Facebook, Skype, etc.).

REFERENCES

[1] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, doi: 10.1145/365153.365168.

[2] K. M. Colby, *Human-Computer Conversation in A Cognitive Therapy Program*. Boston, MA, USA: Springer, 1999, pp. 9–19, doi: 10.1007/978-1-4757-5687-6_3.

[3] B. AbuShawar and E. Atwell, "ALICE chatbot: Trials and outputs," *Computación y Sistemas*, vol. 19, no. 4, Dec. 2015.

[4] A. M. Turing, "I.—Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[5] R. Chatbot. (2008). [Online]. Available: http://brilligunderstanding.com/rosedemo.html

[6] H. Kazi, B. S. Chowdhry, and Z. Memon, "Medchatbot: An umls based chatbot for medical students," *Int. J. Comput. Appl.*, vol. 55, no. 17, pp. 1–5, 2012.

[7] J. Jia, "CSIEC: A computer assisted english learning chatbot based on textual knowledge and reasoning," *Knowl.-Based Syst.*, vol. 22, no. 4, pp. 249–255, May 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705109000045

[8] B. Heller, M. D. Proctor, D. Y.-O.-N. Mah, L. Jewell, and B. Cheung, "Freudbot: An investigation of chatbot technology in distance education," in *Proc. World Conf. Multimedia, Hypermedia Telecommun.*, 2005.

[9] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu, "AliMe chat: A sequence to sequence and rerank based chatbot engine," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2017, pp. 498–503. [Online]. Available: https://www.aclweb.org/anthology/P17-2079

[10] L. Vaira, M. A. Bochicchio, M. Conte, F. M. Casaluci, and A. Melpignano, "MamaBot: A system based on ML and NLP for supporting women and families during pregnancy," in *Proc. 22nd Int. Database Eng. Appl. Symp. IDEAS*, 2018, pp. 273–277.

[11] B. E. V. Comendador, B. M. B. Francisco, J. S. Medenilla, S. M. T. Nacion, and T. B. E. Serac, "Pharmabot: A pediatric generic medicine consultant chatbot," *J. Autom. Control Eng.*, vol. 3, no. 2, pp. 137–140, 2015.

[12] K. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *Proc. 18th IEEE Int. Conf. Mobile Data Manage. (MDM)*, May 2017, pp. 371–375.

[13] I. Divya, P. Ishwarya, and K. Devi, "A self-diagnosis medical chatbot using artificial intelligence," *J. Web Develop. Web Designing*, vol. 3, no. 1, pp. 1–7, 2018.

[14] P. (.-T. Hsu, J. Zhao, K. Liao, T. Liu, and C. Wang, "AllergyBot: A chatbot technology intervention for young adults with food allergies dining out," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst. CHI EA*, 2017, pp. 74–79, doi: 10.1145/3027063.3049270.

[15] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "SuperAgent: A customer service chatbot for E-commerce websites," in *Proc. ACL, Syst. Demonstrations*, 2017, pp. 97–102. [Online]. Available: https://www.microsoft.com/en-us/research/publication/superagent-customer-service-chatbot-e-commerce-websites/

[16] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 3506–3510, doi: 10.1145/3025453.3025496.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, vol. 2, 2013, pp. 3111–3119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999959

[18] H. Al-Zubaide and A. A. Issa, "OntBot: Ontology based chatbot," in *Proc. Int. Symp. Innov. Inf. Commun. Technol.*, Nov. 2011, pp. 7–12.

[19] D. A. Ali and N. Habash, "Botta: An arabic dialect chatbot," in *Proc. COLING*, 2016, pp. 208–212.

[20] B. Hettige and A. Karunananda, "Octopus: A multi agent chatbot," in *Proc. 26th Int. Conf. Comput. Linguistics, Syst. Demonstrations (COLING)*. Osaka, Japan: COLING Organizing Committee, 2016. [Online]. Available: https://www.aclweb.org/anthology/C16-2044

[21] R. Zhao, O. J. Romero, and A. Rudnicky, "SOGO: A social intelligent negotiation dialogue system," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, Nov. 2018, pp. 239–246, doi: 10.1145/3267851.3267880.

[22] C. Segura, A. Palau, J. Luque, M. R. Costa-jussà, and R. E. Banchs, "Chatbol, a chatbot for the spanish 'la liga,'" in *Proc. 9th Int. Workshop Spoken Dialogue Syst. Technol.*, 2018, pp. 319–330.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*. New York, NY, USA: Association for Computing Machinery, 1992, pp. 144–152, doi: 10.1145/130385.130401.

[24] X.-S. Vu, A. Ait-Mlouk, E. Elmroth, and L. Jiang, "Graph-based interactive data federation system for heterogeneous data retrieval and analytics," in *Proc. World Wide Web Conf. WWW*, 2019, pp. 3595–3599, doi: 10.1145/3308558.3314138.

[25] X.-S. Vu and L. Jiang, "Generic multilayer network data analysis with the fusion of content and structure," in *Proc. 20th Int. Conf. Comput. Linguistics Intell. Text Process.*, Apr. 2019, pp. 1–18.

[26] X. Li and D. Roth, "Learning question classifiers," in *Proc. 19th Int. Conf. Comput. linguistics*, 2002, pp. 1–7, doi: 10.3115/1072228.1072378.

[27] R. G. Athreya, A.-C. Ngonga Ngomo, and R. Usbeck, "Enhancing community interactions with data-driven Chatbots–The DBpedia chatbot," in *Proc. Companion The Web Conf. Web Conf. WWW*, 2018, pp. 143–146, doi: 10.1145/3184558.3186964.

[28] S. Keyner, V. Savenkov, and S. Vakulenko, "Open data chatbot," in *The Semantic Web, ESWC 2019 Satellite Events*, P. Hitzler, S. Kirrane, O. Hartig, V. de Boer, M.-E. Vidal, M. Maleshkova, S. Schlobach, K. Hammar, N. Lasierra, S. Stadtmüller, K. Hose, and R. Verborgh, Eds. Cham, Switzerland: Springer, 2019, pp. 111–115.

[29] S. Neumaier, V. Savenkov, and S. Vakulenko, "Talking open data," in *The Semantic Web, ESWC 2017 Satellite Events*, E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, and O. Hartig, Eds. Cham, Switzerland: Springer, 2017, pp. 132–136.

ADDI AIT-MLOUK received the Ph.D. degree in computer science from Cadi Ayyad University, in 2018. He is currently a Postdoctoral Researcher with the Department of Computing Science, Umeå University, Sweden. His research interests include data mining, text mining, machine learning, semantic web, information retrieval, natural language processing, and artificial intelligence. He has collaborated actively with researchers in several disciplines of computer science, particularly sensor networks, smart grid, and VANETs. He has served as a member of conferences and workshops program committees as well as the reviewer for many journals and conferences.

LILI JIANG received the Ph.D. degree in computer science from Lanzhou University, China, in 2012. She was a Research Scientist at NEC Laboratories Europe, Germany, and a Postdoctoral Researcher at the Department of Databases and Information Systems, Max-Planck-Institut für Informatik, Saarbrücken, Germany. She has been dedicating to address academic challenges motivated from real applications by applying the state-of-the-art techniques and exploring novel solutions. She is currently an Associate Professor with the Department of Computing Science, Umeå University, Sweden, and leading the research group of Deep Data Mining. Her research interests include text mining, information retrieval, natural language processing, machine learning, and privacy preservation.