



Retrieval-based Statistical Chatbot in a Scientometric Domain

Victor Lopez-Rodriguez^(iD) and Hector G. Ceballos^(✉)^(iD)

Tecnologico de Monterrey, Monterrey, Mexico

ceballos@tec.mx

<https://www.tec.mx/>

Abstract. The scope of this research work is to integrate a statistical ontology model of scientometric indicators in a chatbot. Building a chatbot requires the use of Natural Language Processing (NLP) as a capability for recognizing users' intent and extracting entities from users' questions. We proposed a method for recognizing the requested indicator and transforming the question expressed in natural language into a query to the semantic model. The chatbot and the ontology model represent a novel framework that can answer questions about Scientometric Indicators. The chatbot is evaluated in terms of Goal Completion Rate (GCR). It measures how many questions the chatbot answered correctly and identifies intent and entity extraction correctly. The second evaluation approach of the chatbot is a survey that focuses on usability, the strictness of language variations, chatbot comprehension, correlation in chatbot responses, and user satisfaction.

Keywords: Chatbot · Statistical ontology · Natural language processing · Scientometric indicators

1 Introduction

A chatbot is a machine conversation system that interacts with human users via natural conversational language [1]. Chatbots are increasing their appearance in several environments related to software interaction with humans. This software is used to perform tasks such as quickly responding to users, informing them, helping them purchase products, and providing better service to customers [2]. Common chatbots' applications are Frequently Asked Questions (FAQ), Customer Support, and helping users obtain fast answers to their questions. The chatbot creates strength in the methodology by implementing several Natural Language Processing techniques that allow the user to feel a one-to-one conversation. One of the best advantages of using a chatbot is creating context in a conversation. Comparing the chatbot against other alternatives such as Online Query or Dashboards, the conversation agent can use the conversation history, including questions and answers, to create a context and have a fluent conversation with correct responses.

Supported by organization x.

The chatbot domain is about Scientometric Indicators. The term scientometrics was coined by Vassily V. Nalimov in the 1960s and referred to the science of measuring and analyzing science, such as a discipline's structure, growth, change, and interrelations [3]. Scientometric analysis studies the quantitative areas of the process of science, science policy, and communication in science by having a focus on the measure of authors, articles, journals, and institutions by understanding citations related to them [4].

Reviewing the state of the art, we found a relevant research work that present the use of a chatbot for solving a specific task. Baby [8] presented a chatbot using Natural Language Processing techniques to understand and extract key information on user requests. The chatbot was connected to a web application for home automation and performed several tasks such as fan or light controlling and other electrical appliances. In our research work, a chatbot approach is to fully understand the request and create a context to have better insight for answers.

The main research question of this work relies on how a retrieval-based chatbot can be used along with an ontology model as a knowledge-base to answer questions in natural language in the Scientometric Indicator domain. We stated the following specific research questions to design the methodology: How to fully understand users' intents related to scientometric indicators? Which important keywords does the chatbot need to recognize to perform the tasks? If there is not enough information to answer the question, how will the chatbot converse to create a context to gather the needed data?

The objective of this research work is apply an ontology model that allows the chatbot to understand and classify the user's input to extract the correct information and answer the question correctly. The goal is to propose and design a scalable chatbot that can be used in the future in other academic areas or industries.

The chatbot will be validated in terms of completeness. Completeness refers to a specific set of indicators in which the chatbot recognizes and answers correctly. Time validation is a subject of study in performance; unfortunately, previous work does not exist that we can use to compare the new development in this research work.

This research work consists of 4 Sections including the introduction. Methodology is presented in Sect. 2. In Sect. 3 results are shown along its discussion. Finally, in Sect. 4, we present conclusions and future work.

2 Methodology

The strategy that the chatbot follows is a retrieval-based approach. For this research work, we will extend the work proposed in [5] which a semantic modeling of scientometric indicators using the ontology Statistical Data and Metadata (SDMX) is proposed. This ontology contains a set of scientometric indicators

stored in a graph database in Neo4j. It is an updatable ontology model and will provide a quick access for retrieving information in our proposed chatbot.

Our chatbot is classified in the closed domain of Scientometric Indicators. It is a task-based chatbot whose main task is to answer questions related to the mentioned domain. In order to answer the questions, the chatbot will follow a retrieval-based strategy that will use the ontology model to dig into data and return values in the form of answers.

2.1 Natural Language Processing

This section of the methodology will analyze the natural language questions that the users may ask to the Chatbot. As one of the initial steps, some frequently asked questions were retrieved in order to understand their characteristics and patterns.

After we analyze this information, we decided to define a classification that will allow us to detect the intention of the question and prepare the process to answer it. The complexity of questions depends on the number of indicator values recovered and the type of calculations applied for generating the answer. Depending on the complexity of the question, the categories are described as follow:

- Low Complexity: Simple questions with a direct answer, i.e. a single indicator value is required for answering.
- Medium Complexity: Questions that require calculations such as sum, average, difference or further analysis, i.e. two or more indicator values are recovered and combined for providing a new value.
- High Complexity: Questions that require machine learning models for prediction, i.e. multiple indicator values are recovered and a series of expected values are generated.

We also add the greeting and none intent. Greeting intent will be used to identify when the user wants to start a conversation in the chatbot and the None intent is empty at purpose because we can identify questions out of the domain.

2.2 Intent Classification and Entity Extraction

In this step of the methodology we will identify the intention of the question and extract relevant information from it to be able to answer it. An intent can be represented as a task or action the user wants to perform, it can be considered as a purpose or goal. An entity can be defined as words or phrases inside the utterance that describe important information of the intent. We proposed the following entities:

- Indicador: Entity created for finding Scientometric Indicator names.
- Lugar: Entity created for finding a school, set of schools, or referring to Tecnológico de Monterrey that considers all the schools.

- **Objeto:** Entity created for obtaining the “thing” that we are looking for in medium complexity intentions.
- **Tiempo:** Entity created for finding years or time related words.
- **Tipo Indicador:** Entity created for finding if the Scientometric Indicator refers to annual or quinquennial type.

2.3 Data Labeling

The next step is to label entities in all the utterances in each intent type as shown in Table 1. With the intent classification and entities labeling, we are ready for training the model.

Table 1. Intent and Entities labeling

Intent	Utterance examples	Indicador	Lugar	Objeto	Tiempo	Tipo indicador
Greeting	Hola	NA	NA	NA	NA	NA
Low Complexity	Cuántos SNIS hay en el Tec en el 2019?	SNIS	Tec	NA	2019	NA
Medium Complexity	Cuál ha sido el año con el menor número de SNIS?	SNIS	NA	Año	NA	NA
High Complexity	Cuál citas habrá en el 2025?	citas	NA	citas	2025	NA
None	Empty	NA	NA	NA	NA	NA

2.4 Model Training

We choose to use an Artificial Intelligence service from Azure Cognitive Service called Language Understanding (Luis). This service applies custom machine learning intelligence to user’s conventional natural language text to predict the meaning or pull out relevant information. One of the reasons for choosing this approach is the integration with the Chatbot framework of Azure.

The next step of the process is to train the model. Training is performed in Luis AI service and is done iteratively. We start by randomly selecting 5 utterances with different classification of intents and already having the entities labeling done in the previous step. We train the model and make some tests to observe the accuracy of correct identification and extraction. We continue doing this step iteratively until we reach the 50 utterances including several variations to the questions to achieve a better result.

2.5 Scientometric Indicator Identification

In this methodology process, we will use the intent identification and entities extraction of the natural language question to identify the Scientometric Indicator that the user wants an answer about. The first part is to obtain descriptive information of the Scientometric Indicators and we will obtain this by querying the ontology model as shown in Fig. 1.



Fig. 1. Query for obtaining Scientometric Indicators Dataset labels and description.

By matching all the nodes of type dataset and return all the rdf:labels and comments of the dataset that will allow us to make a knowledge base of Indicators for the Chatbot. The result of the query will be transformed from a json data type to a dictionary to be able to use it in a further process. The knowledge base of True Scientometric Indicators is shown in the following list and it contains the description of the Indicator and relevant tags extracted from the label node property.

- annual publications scopus-tec dataset: ['publication', 'annual']
- annual school publications dataset: ['school', 'publication', 'annual']
- quinquennial school publication dataset: ['quinquennial', 'school', 'publication']
- quinquennial school cites dataset: ['quinquennial', 'school', 'cite']
- annual school document cites dataset: ['cite', 'school', 'annual']
- researchers dataset: ['researcher', 'annual']
- posdocs dataset: ['posdoc', 'annual']
- quinquennial publications dataset: ['quinquennial', 'publication']
- quinquennial cites dataset: ['quinquennial', 'cite']
- document cites dataset: ['quinquennial', 'cite']

Suppose we have the following natural language question: How many publications were made in the quinquennium ending in 2021?, by sending this question to Luis AI service, it will return the following response.

- topIntent: Low Complexity
- entities: indicador: ‘publicaciones’, tipo indicador: ‘quinquennales’, tiempo: ‘2021’, Lugar: ”

In order to identify the correct Scientometric Indicator to which the natural language question is referring, we will work with both sets. Our approach for a correct identification is to use the set theory.

- Equal Sets: When elements (labels) are the same members of True Scientometric Indicator and Tags Sets. Also called super sets.
- Proper Subset: When elements (labels) from Tag Set are included in the True Scientometric Indicator Set elements but still have other elements missing to be a Super Set.
- None equal Sets: Both sets have difference elements.

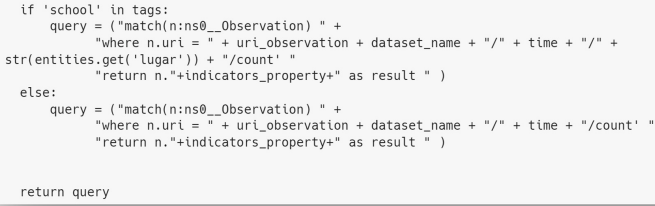
The result of matching the indicator of the mentioned question is having an equal set with the quinquennial publications dataset indicator.

2.6 Natural Language Transformation into Cypher Query

In this last step of the methodology, we will use the intent identification provided by Luis AI service. Depending on the complexity intention, there is an structured query ready with some parameters to fulfill with the entities extracted from the natural language question. Our approach for this chatbot will only focus in the low complexity intention. We define the following parameters to fulfill the query structure:

- uri_observation: Uses the following URI (<https://www.tec.mx/ontos/observation/>) defined in our ontology model for querying Scientometric Indicator observations.
- dataset_name: This value is obtained from the True Scientometric Indicator Set in which the descriptive value is selected by matching with the Exact Indicator key.
- time: Numerical value obtained after comparing with Time Knowledge-base, if the indicator is of type annual we will only use a year. However, if it is quinquennial, we will use the range period of year-(year-5).
- school: If the label school is in our tags, the structure changes and we add the condition by obtaining the value of the entity Lugar.
- indicators_property: This value is obtained from the Measure knowledge base by matching the Exact Indicator key.

After filling up all the values of our parameters, we run the following query in Fig. 2.



```

if 'school' in tags:
    query = ("match(n:ns@_Observation) " +
            "where n.uri = " + uri_observation + dataset_name + "/" + time + "/" +
            str(entities.get('lugar')) + "/count' "
            "return n."+indicators_property+" as result " )
else:
    query = ("match(n:ns@_Observation) " +
            "where n.uri = " + uri_observation + dataset_name + "/" + time + "/count' "
            "return n."+indicators_property+" as result " )

return query

```

Fig. 2. Natural language question into cypher query.

2.7 Chatbot Deployment

The first step to start building the Chatbot for Scientometric Indicator is to use a tool provided from the Azure Bot Service called Bot Framework Composer. The chatbot runs through a flow diagram in which two main concepts are Dialog and Triggers. Dialogs are a central concept in the SDK, providing ways to manage a long-running conversation with the user. It performs a task that can represent part of or a complete conversational thread, and it can span just one turn or many and span a short or long time-period [6]. Dialog triggers handle dialog specific events that are related to the life-cycle of the dialog [7].

We built and integrated our Scientometric Indicator Chatbot with the ontology model in this section. We built a Scientometric Indicator API (SI API) that enables communication between the Chatbot Framework in Azure and the ontology model in Neo4j. An architectural diagram of the main components is shown in Fig. 3.

The Scientometric Indicator API was created in python language using the flask framework for a faster development. The goal of the API is to connect the several components such as the ontology model in Neo4j, the chatbot framework and the Luis AI service for intent identification and entities extraction.

The process of hosting the chatbot for the testing period was achieved using Google Sites for sharing the Chatbot with the users.

3 Results and Discussion

In this section we shown the results by testing the chatbot with users from the Research Office at Tecnológico de Monterrey. For this evaluation we host the Scientometric Indicator Chatbot in the web and make it available for the users for testing. The users had 3 weeks for testing several natural language questions to the Chatbot and at the end evaluate it with a survey that will contain the following questions with answers from 1 to 5. 1 meaning Few Knowledge or Deficient and 5 meaning Expert or Excellent depending on the type of question.

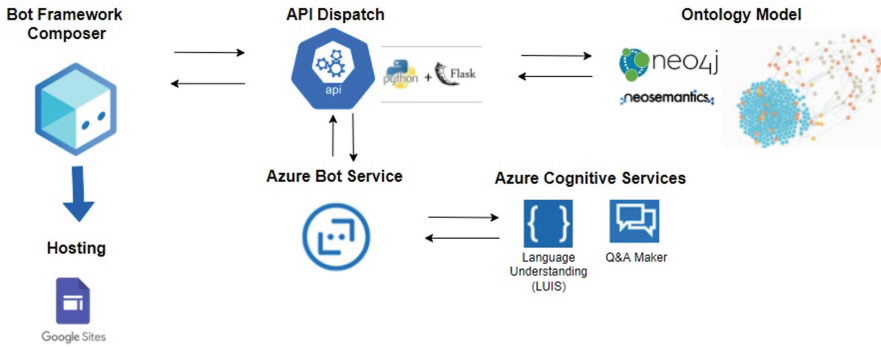


Fig. 3. Architectural diagram

- User's email
- User's knowledge about Scientometric Indicators
- Usability: Ease of use and time required for the Chatbot to answer the questions.
- Strictness: Ability from the chatbot to understand language variations.
- Comprehension: Ability from the chatbot to understand the question and answer respect the relevant information.
- Correlation: Relevance of the questions according to the context of the question.
- Satisfaction: User's feeling with the Chatbot and it's future.
- Comments or Feedback

Along with this survey, we will log all the users' interactions with the chatbot. With this logs we will evaluate the following metrics:

- Structure of the Conversation: Number of users and Total Conversations.
- Goal Completion Rate: Number of times chatbot answered correctly, number of correct intention detected, correct indicator detected and correct entity extraction.
- Bot Response Time: Comparison between the time took to answer a question of the chatbot and the actual process.

3.1 Goal Completion Rate

A total of 11 users participated in this test in which they made 35 different questions about Scientometric Indicators to the Chatbot. Whenever the users asked the chatbot, relevant information such as the date-time, user, top intent, exact indicator, comments, answer, and list of possible indicators were retrieved and stored in the ontology model. This section will evaluate the Goal Completion Rate (GCR) in terms of answers, intent detection, entity extraction, and scientometric indicators identification.

The first evaluation is the GCR of the correct intention detection. Intent detection is a great resource for correctly retrieving the scientometric indicator value. In Table 2 we can observe that 21 questions were identified as Low Complexity intention. Users made 6 questions of Medium complexity and finally, 8 questions were identified with the None and Greeting intentions. The GCR evaluates how many intentions were correctly identified by the chatbot. In Fig. 4 we can observe that 25 intentions, that represent 71.42% from the total questions, were identified correctly.

Table 2. Intent identification

	Intent	Number of questions	Correct intent
1	Low Complexity	21	21
2	Medium Complexity	6	3
3	Greeting	7	0
4	None	1	1

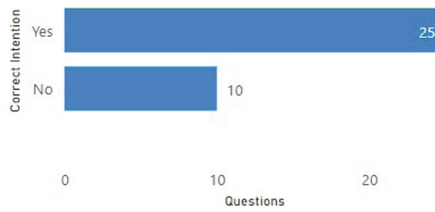


Fig. 4. Goal completion rate: correct intention

This result means that the training helped a lot in terms of variances in the questions due to complexity. Obtaining this result allowed the chatbot to understand what process to follow to answer the question correctly.

The second evaluation consists of assessing the GCR of the correct identification of scientometric indicators. In Fig. 5 we can observe that the chatbot identified the scientometric indicator correctly in 17 questions made by the users. This outcome represents 48.6% of the total questions.

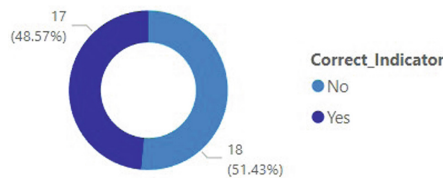


Fig. 5. Goal completion rate: correct indicator

The next evaluation is about the GCR from entity extraction. Entity extraction allows us to find relevant data for retrieving the scientometric indicator value more precisely. In Table 3 we can observe that 23 questions entities were extracted correctly and it represents 65.71% from all the question.

Table 3. Entity extraction

	Number of questions	Correct entity
1	23	Yes
2	12	No

We can state that the chatbot correctly identifies these entities in 65.71% of the questions due to differences in asking questions and missing required data.

In order to evaluate the GCR for correct answers from the chatbot to the questions, we need to establish the following categories.

- 1: Answered correctly with the required value
- 2: Answered incorrectly with a value or could not understand the question due to indicator matching, intent identification, or entity extraction.
- 3: The chatbot understood the question correctly but could not answer the question because data was not available in the ontology. In other words, the value for that time (year or quinquennium) is not stored in the ontology.

In Table 4 we can observe that the chatbot answered 14.3% of the questions correctly with a proper value from the ontology. In the other case, the chatbot could not answer 51.4% of the questions because it could not understand the question properly due to indicator matching or entity extraction. In the last case, the chatbot answered 34.3% correctly because it understood the question, but unfortunately, the value in the period was not uploaded for this test.

With the previous results, we decided to group categories 1 and 3 because the chatbot understood the question and answered correctly according to the values stored in our ontology due to the problem nature. The results are shown in Fig. 6 and state that the chatbot answered correctly 48.6% of the questions from the test evaluation while 51.4% were answered incorrectly.

Table 4. Goal completion rate: correct answer

ID	Number of questions	Percentage
1	5	14.3%
2	18	51.4%
3	12	34.3%

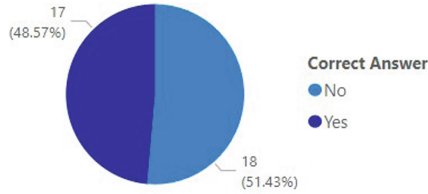


Fig. 6. Goal completion rate: correct answer

From this evaluation approach, we can state that every time the chatbot correctly identified the scientometric indicator, it answered the question correctly. It is essential to state that the chatbot only had one round of training and obtained reasonable goal completion rates.

Our research highlights the approach of defining three types of complexities for intention identification by obtaining a good result in the evaluation. However, this only represents a part of the process of interpreting the question correctly. There can be several issues like not extracting the correct entities and not filtering the desired value or not having the value stored in the model. Issues can be solved by adding more training steps to the model.

Bot Response Time. We extracted the chatbot's time to answer the questions and calculated the average from the testing log data. The chatbot took approximately 1.5s to answer a question of scientometric indicators, and the actual process takes approximately 10s to retrieve the answer.

3.2 Survey Evaluation

This evaluation consists of a survey in which users submitted their responses after testing the chatbot. The survey contained questions about metrics such as user confidence in their knowledge about scientometric indicators at Tecnológico de Monterrey, chatbot usability, strictness in language variations, chatbot comprehension, correlated replies from the chatbot, and users' satisfaction over the chatbot. The response was available from 1 to 5, 1 meaning low, and 5 refers to a high level. We had 11 users from the Research Office at Tecnológico de Monterrey who participated in the survey, and the replies are shown in Table 5.

Table 5. Survey data

User	User confidence	Usability	Strictness	Comprehension	Correlation	Satisfaction
1	4	4	1	1	1	1
2	5	3	1	1	1	1
3	4	4	3	3	3	4
4	4	4	1	1	1	1
5	4	1	1	1	1	3
6	5	4	1	1	1	1
7	3	4	3	3	3	3
8	4	4	3	4	4	4
Avg	4.2	3.5	1.75	1.87	1.87	2.25

This result was expected because the chatbot was trained with only a language format and the questions made during the testing period had different language variations in the questions. We decided only to briefly explain the chatbot and not provide sample questions not to produce bias in the users. Users tested the chatbot without any experience, which provoked several misunderstandings in the way of asking questions.

Finally, we also consider that these results would have improved if more training steps were added to the model to increase the chatbot's capacity to understand questions. We found much feedback from this evaluation approach that will allow us to improve the chatbot.

4 Conclusion

The main research questions stated in this research was proven correct. This research demonstrated that a chatbot could be used along with a statistical ontology model that extends SDMX to correctly answer any given questions about Scientometric Indicators. We can state that the specific objectives have been met, and the research questions were answered during the development of this research work.

The Natural Language Process allowed us to answer the research question of how the chatbot talks with the user to create a context for gathering data when there is not enough information to answer the question. The chatbot needs to identify the question's intention and extract the entities needed to understand the question and provide a correct response. It allows us to answer the research question on which essential keywords the chatbot must recognize to perform the task.

The chatbot answered correctly in almost 50% of the questions made to it. We can conclude that we met the goal of applying an ontology model that allowed the chatbot to understand and classify the user's input to extract the correct information of the question and provide a correct answer. The chatbot development knowledge was designed to create context during the conversation where information was needed to understand the user's question. It allows us to

meet the objective of proposing and designing a scalable chatbot that creates a context for missing information for answering a question that can be used in other academic areas or industries.

Many different features and developments, also considered opportunity areas, have been left for the future. This research work can be considered the initial step for using the statistical ontology model and the chatbot in the Research Office at Tecnológico de Monterrey. Future work will concern with the following aspects:

- It will be interesting to add medium and high complexity knowledge to the chatbot in the future. Having a feature of using aggregation functions in time intervals and using stored information to forecast the number of cites in a particular year using machine learning models can lead to a very intelligent chatbot.
- Improve the training in Luis AI service to achieve higher results in Goal Completion Rate in answering the questions correctly, correctly identifying scientometric indicators and correct entity extraction.
- In the chatbot deployment, we can improve the usability by using thumbnail cards during the conversation to guide the user in building the question according to the knowledge, context, and stored data in the ontology model. This feature will help in having fewer questions with missing entities.

Acknowledgements. We thank Tecnológico de Monterrey and CONACyT for the financial support.

References

1. Shawar, B.A., Atwell, E.S.: Using corpora in machine-learning chatbot systems. *Int. J. Corpus Linguist.* **10**(4), 489–516 (2005)
2. Albayrak, N., Özdemir, A., Zeydan, E.: An overview of artificial intelligence based chatbots and an example chatbot application on Proceedings, pp. 1–4. IEEE(2018)
3. Hood, W., Wilson, C.: The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics* **52**(2), 291 (2001)
4. Zakka, W., Lim, N.H.A., Chau, M.: A scientometric review of geopolymer concrete. *J. Clean. Prod.* **280**, 124353 (2021)
5. Lopez-Rodriguez, V., Ceballos, H.: Modeling scientometric indicators using a statistical data ontology. *J. Big Data* **9**(1), 1–17 (2022)
6. Dialogs library. <http://docs.microsoft.com/en-us/azure/bot-service/bot-builder-concept-dialog?view=azure-bot-service-4.0>. Accessed 27 Mar 2020
7. Events and triggers in adaptive dialogs - reference guide. <http://docs.microsoft.com/en-us/azure/bot-service/adaptive-dialog/adaptive-dialog-prebuilt-triggers?view=azure-bot-service-4.0>. Accessed 27 Mar 2020
8. Baby, C., Khan, F., Swathi, J.N.: Home automation using IoT and a chatbot using natural language processing. In: 2017 Innovations in Power and Advanced Computing Technologies (i-PACT) on Proceedings, pp. 1–6. IEEE (2017)
9. Chen, Z., Lu, Y., Nieminen, M., Lucero, A.: Creating a chatbot for and with migrants: chatbot personality drives co-design activities. In: 2020 ACM Designing Interactive Systems Conference on Proceedings, Eindhoven, Netherlands, pp. 219–230. Association for Computing Machinery (2020)