KGBot:A Knowledge Graph Based Chatbot Utilizing Linked Data

1st Shazil Ahmed
FAST NUCES
i191910@nu.edu.pk

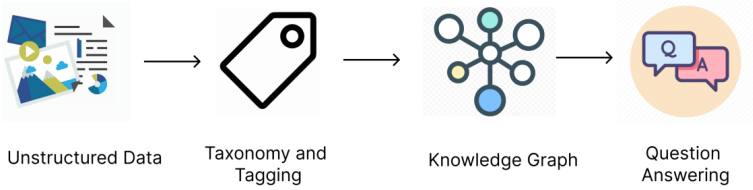
2nd Zaafir Rizwan
FAST NUCES
i191723@nu.edu.pk

3rd Zohaib Bashir
FAST NUCES
i191669@nu.edu.pk

***Abstract*—Ever since Google embraced knowledge graphs in 2012, there has been a steady increase in the progress of the semantic web. A huge amount of data specifically structured data has become available on the internet in the form of (KBs) Knowledge Bases. Making the data within these knowledge bases accessible to the general public is one of the main objectives of chatbots which employ the use of linked data. Creating a chatbot over linked data raises various difficult challenges such as user query understanding, support of multiple different knowledge bases etc. To handle these challenges we propose a machine learning approach based on intent classification and NLU (Natural Language Understanding) to understand user intents and generate SPARQL queries. The system can be extended with a new domain on-demand, multiple knowledge bases, flexible and allows execution of different tasks for an extensive range of topics. We show through evaluation that our chatbot is more effective in understanding user utterances with domain-specific entities.**

. ***Index Terms*—Knowledge Based systems, Chatbots, Linked Data, Natural Languages, Semantic Web, Machine Learning, Information Retrieval, Natural Language processing, Query Processing, Structured Data, Multiple Knowledge Base Support, SPARQL**

I. Problem Statement



Since the world and the technology are evolving on a day by day basis, a guide is required which will help us in understanding the use of the technology and our surroundings. With the advancement of the web specifically the Semantic Web, A lot of data that being structured data is present in the form of Knowledge Bases. Building a Chatbot on linked data has many challenges such as query understanding from the user's side, compilation of multiple sources of data etc. In the past several years, the big tech giants such as Google, Apple, Amazon all have invested in AI and developed several chatbots among them are Siri, Microsoft Cortana and Alexa. In the context of linked data the main purpose of a chatbot is to retrieve useful and relevant information from either one or multiple sources of information such as KBs (Knowledge Base) by using NLU and semantic web technologies. Recently, with the growth development of linked data, increasing progress on chatbots has been seen in research and industry.

II. Introduction

Numerous chatbot frameworks have been proposed, however they required a ton of training data which was unavailable and was expensive to create. As of recent times due to the Semantic Web and the growth of linked data has increased the progress of chatbots both in the research and industry fields. However there are still many challenges such as including user queries understanding, multiple knowledge base, intent classification, and analytical queries understanding. In this research we propose a chatbot (KGBot) that is able to address some of the challenges listed above. Some of our contributions are as follows.

- We built and designed KGBot.
- We build a classifier for intent classification using a machine learning model
- Insuring scalability, KGBot is flexible by adding other sources of information i.e. other knowledge bases

A. Motivation

As of 2022, the Chatbots of the world are not as intelligent as they can be. The industry is more focused on other groundbreaking technologies such as Artificial Intelligence, Edge Computing and quantum computing which have their own place in the industry. The Chatbots that do exist out there are made for a specific field or for a specific organization. The Chatbot can only function within those boundaries. A normal Chatbot is also not able to answer informal answers. The problem with existing question answering systems is that they answer the query/questions of the user independently, adding redundancy for the user to repeat their previous question even if a follow-up was already asked. Normal Chatbots are limited to either keywords or fixed responses as they are

made in this way. Normal Chatbots don't work in real time meaning that the information retrieval of a normal Chatbot is already predetermined and it does not have an answer for any information that was made available after the creation of the Chatbot.

B. **Background**

User data has existed for a long time. Before the age of computers it was stored in files but with the revolution of computers it was then stored in file systems which then was transformed to databases to store the information. One caveat of databases is that data is not linked to one another. For example a person's name has no relation to that same person's age. User has to manually make the connection via queries which only exist for those queries. For this purpose a knowledge model was made that is used to make connections between the various entities. This knowledge model is the heart of a Knowledge graph. A Knowledge Graph is used to put data in context via linking and semantic metadata which provides a framework for data unification, data integration and data sharing. The information is usually stored in a graph database and visualized as a graph structure which is the reasoning for the term knowledge "graph." A Knowledge Graph is made up of three main components those being nodes, edges, and labels. Any object, place, or person can be a node. An edge defines the relationship between the nodes. Fig

2. Where A represents the subject, B represents the predicate and C represents the object. Take the example of a car. A Car is a Vehicle, where A represents Car, B represents is a relation and C represents Vehicle.

Knowledge graph are typically made up of datasets from various resources which differ in their structure. Schemas and context are used together to provide structure to diverse the data. These components help in distinguishing words which have multiple meanings. Nowadays Knowledge Graph are being used everywhere such as they are used in recommendation systems, cyber security, predictive maintenance etc. The big tech giants also employ the use of KG such as Google, Amazon, and Facebook etc.

A Chatbot is a computer program which is used for automation purposes. A Chatbot employs the use of Artificial Intelligence (AI) and Natural Language Processing (NLP) to understand user questions and automate the responses. Chatbots are also known as virtual agents or virtual assistants. They allow the user to retrieve the information they need without the use of a human intervention. A Chatbot can be queried via text input, audio input or both. Chatbots are used in various ways such as they are used for

- Providing Directions
- Receiving customer support from a specific company (mobile, TV etc.)
- Finding Local shops

Chatbots have also alleviated the burden on many companies which are short on manpower and costs as a Chatbot is active 24/7 whereas a human will require breaks. A Chatbot also reduces cost and improves user engagement. A Chatbot makes it easier to collect data and improve the experience.

III. Related Work

A. **Knowledge Graph Chatbot using Linked Data:**

In this paper [1] they list the challenges of creating a Knowledge Graph based Chatbot. Due to the rapid progress of the Semantic web, a large amount of data specifically structured data has become available in the form of KBs. A Chatbot that depends on linked data has many challenges such as user queries understanding, multiple KB support and a multilingual aspect as well. They employed the use of a machine learning approach based on intent classification and NLU to understand user intents and generate SPARQL queries. The proposed approach is based on a modular approach which takes advantage of semantic web techniques, KG and machine learning. The proposed Chatbot can handle different tasks (such as FAQs, Analytical queries etc.), gathering information from multiple sources and presenting them in the form of a knowledge card. The proposed Chatbot is developed using Flask framework and can run on a standalone or distributed mode to improve response time of information retrieval.

B. **Hybrid Chatbot:**

The paper [2] talks about incorporating both Knowledgebase and question answering approaches. In the paper it is explained that the Q n A portion is done via machine learning techniques and Knowledge base portion is done via a metagraph model which requires hybridization of two approaches. For Q n A machine learning techniques used are tokenization, lemmatization and vertorization for TF-IDF processing all of which are components of the library NLTK. The vectorized question is passed to the Q n A processing module whereas for knowledgebase the preprocessing is for concept recognition. This idea of concept recognition is based on NLTK sentence parsing. The knowledgebase processing module employs the use of the metagraph approach. Metagraph is a kind of complex graph model aimed for hierarchical graph description. It simplifies N-ary relationship representation and complex contexts description. The metagraph rule agents are used for KGB Q n A and an active dialogue with the user.

C. **Interactive Q n A:**

In this paper [3] it is explained how a QA (Question Answering) System is defined as a QA system that supports a series of exchanges between users and systesms to clarify user intent and to enable follow-ups. In this paper a technique called Adobot is used for designing and building What and How questions answering systems. The paper's main contribution is providing a solution that integrates an ontology-based KG focusing on What and How questions as well as user context. The ontology based Knowledge Model is aimed to process

knowledge at the semantic level to improve accuracy of user queries. Adobot is named after the case study of Adobe Photoshop. It includes different knowledge models, architecture and its implementation.

D. Building Chatbot via Large Scale Knowledge Base:

In this paper [4] they list the lessons and the challenges learnt from building a large scale virtual assistant for understanding and responding to equipment related complaints. They provide an alternative scalable framework for extracting knowledge from short text and identify entities in user utterances. One of the benefits of the purposed KB framework is facilitating the development and deployment of intelligent assistants for various industrial AI scenarios. Another benefit of this paper was that it improves existing maintenance solutions through better processing of user complaint text.

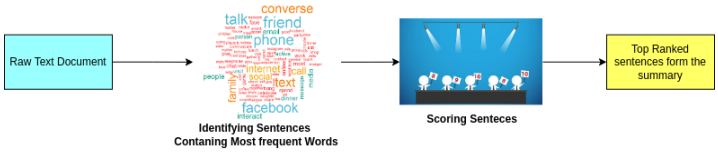
IV. Our Approach

Our proposed approach combines various Natural Language Processing techniques in order to solve the aforementioned problem. We will specifically focus on Text Summarization, Coreference Resolution, Named Entity Linking, and Relationship Extraction. It is essential to mention that the solution cannot be limited to only these techniques; instead various other algorithms and procedures can be used. However, our solution will focus mainly on the above mentioned techniques.

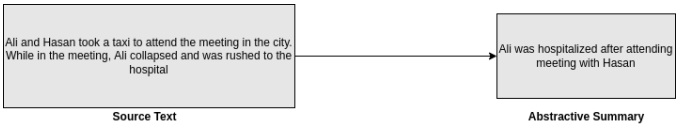
A. Brief Description of the techniques:

This section contains a short description that delineates how each of the techniques work and what can be achieved using them. The next section will specifically talk about how these approaches can be used to solve our problem.

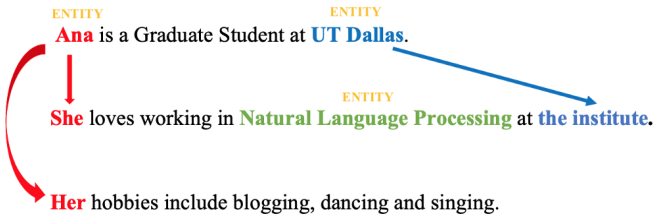
1) Text Summarization: As the name suggests, text summarization is an NLP technique used to summarize the text in a document. Mainly there are two types of summarization techniques in NLP, extractive and abstractive. In extractive summarization, the technique focuses on extracting important sentences from the document, combining them and formulating the summary. It is to note here that the sentences used to formulate the summary are exactly the same as in the original document. Some of the most popular extractive summarization techniques include Text Rank and Luhn algorithm. The Text Rank algorithm works by identifying words that occur most frequently in document and the identify sentences that contain these frequent words. According to this, a numerical score is assigned to each sentence and in the end 'n' numbers of top ranked sentences are picked and used to formulate the summary. ' n ' is a parameter specified by the user. The Luhn summarization algorithm works based on the TF-IDF (Term Document Inverse Document Frequency) technique. It is preferred when neither the highest frequency nor the lowest frequency words are important. Based on this it assigns score to sentences and top ranked sentences are picked for summarization. Extractive text summarization is summed up in the flowchart below. Note that the sentences picked to formulate the final summary remain unchanged.



The other summarization technique in Natural Language Processing is abstractive text summarization. Mainly, it differs from extractive in the sense that it does not pick the highest ranked sentences as it is from the document, rather it generates new sentences that best represent and summarize the text in the document. The state of the art technique used to implement abstractive summarization is using Transformer models fine-tuned specifically on a summarization dataset. Some of the most famous transformers used are BERT, Huggingface, GPT and T5. An example of abstractive text summarization is given below where it is observed that the document information is encapsulated in a new sentence which was not previously present in the doc.



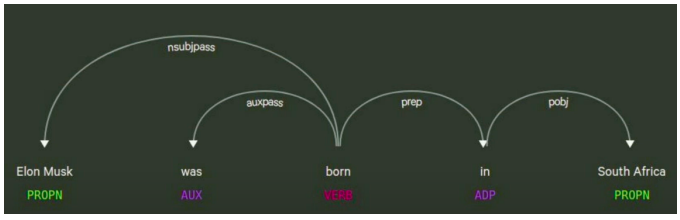
2) Coreference Resolution: Coreference Resolution is a Natural Language Processing technique which aims to replace pronouns in the text document (e.g. he, she , it) with the proper nouns that they are referring to. This is essential as it makes the text more comprehensible for our machines. This is because we humans can understand what a pronoun in the middle or at the end of a paragraph is referring to, but for computers, this is not possible. Therefore, for the sake of accurate information retrieval, it is significant that we remove these pronoun/noun ambiguities from a text document.



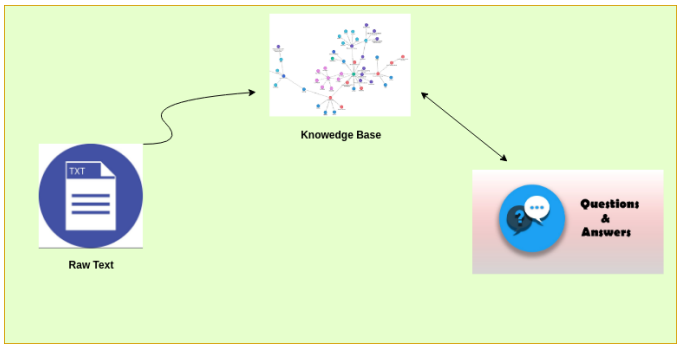
Coreference Resolution mainly works by combining two sub-tasks i.e Mention Detection and Mention Clustering. Mention Detection aims at finding all the candidate spans referring to some entities. This includes all the proper nouns (Ali, Sara, Table) and the pronouns (he, she, it). Then Mention Clustering is applied which aims to identify which of mentions are referring to the same entity. It then combines them together these mentions and replaces the pronoun with its associated noun, hence simplifying the text document.

3) Named Entity Linking (NER): Though coreference resolution replaces pronouns with the associated nouns, it does not help when the same noun is referred to with different names. For example, a person might be referred to by his full name at the start of the document and then somewhere in the middle he might be referred by his last name only. He can also be referred by a designation that although a human can instantly understand that it refers to him but for a machine this becomes difficult. Hence this is where Named Entity Linking (NER) technique of Natural Language Processing is utilized in order to map these different entities referring to the same noun as one. It can be seen in the example below that 'Elon Musk', 'Elon', and 'Musk' refer to the same person. Hence NER helps in removing this ambiguity.

4) Relationship Extraction: Relationship extraction is a Natural Language Processing technique used to extract the semantics embedded in a sentence. The go to method to perform this task is the Part Of Speech (POS) technique. POS assigns a part of speech tag to each word in the corpus. It uses various techniques, some very common occurring words are directly assigned a POS whereas in other situations advanced techniques such as markov chains are used which essentially look back at the previous words and their POS before assigning the POS to the current word.



5) Solving our problem using above mentioned NLP techniqu...
The problem that we are trying to tackle is generation of a knowledge graph from a raw text document and then using the knowledge graph as a knowledge base for a question / answering system.



In the pipeline shown above, the first part which requires conversion of raw text into a knowledge graph is where most of the techniques that we explained in the previous section are going to be utilized. The algorithm for this part is as follows

Algorithm

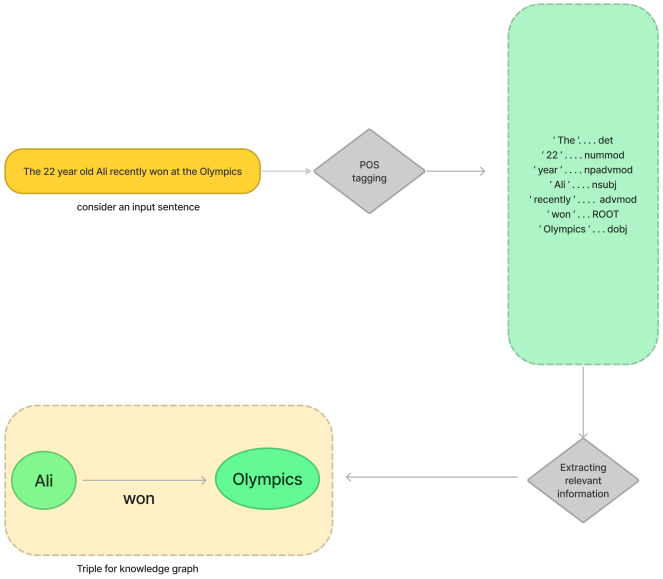
```
function  TextToGraph (  raw tex document  -----> Doc  )

        Doc1  =   CoreferenceResolution(Doc)
        Doc2  =   NamedEntityLinking(Doc1)
        Doc3  =   ExtractiveTextSummarization(Doc2)

        for each sentence in Doc3
              Doc4  +=   EntityExtraction (  POSTagging(sentence)  )


        return    Doc4
```

We start with the raw text document (Doc). We first perform Coreference Resolution(on Doc) in order to remove ambiguities regarding nouns and pronouns. (Result Doc1) We then perform Named Entity Linking(on Doc1) in order to remove ambiguities pertaining to proper nouns being referred with multiple different words. (Result Doc2). We then perform Text Summarization(on Doc2) in order to summarize the text and extract only important, relevant information that would act as solid facts. It is to note that we opt for extractive text summarization as we want to pick the sentences from the document as it is rather than formulate new sentences that might contain unnecessary words that would lead to increase ambiguity. (Result Doc3). Assuming that now each sentence contains the required necessary information, we tokenize the document on the basis of sentences. We then perform, for each sentence, PartOfSpeechTagging followed by Entity Extraction in order to extract the required subject predicate object from the sentence which form the basis triple structure for our knowledge graph. Doc4 now contains the triples for our knowledge graph. This part of the algorithm is illustrated in the image below.



V. Evaluation and Experiments

In order to experiment our approach, we propose to first start with a straightforward text document in which we state some straight forward facts about a person. The language to be used in this initial text document is plain and simple. We shall avoid using complicated sentences and try not to convey a piece of information using too many sentences. In short, our aim here is to keep the text unambiguous and easy for our machine to interpret and extract facts from. We hope to achieve fairly decent results from this dataset because of its simplicity and clarity. [Text 1]
Next we aim to increase the difficulty level of the language in the text document. We aim to pick a text involving fairly routine language used, with various punctuations and a difficult grammar and vocabulary. [Text 2]

VI. Results

We obtained mixed results after passing the text documents through our proposed pipeline. The results were relatively better for Text 1 which had simple and plain language. We were able to extract more than 90 percent of the information correctly and were able to answer the questions, suggesting that our technique works very well where the text document is not large and doesnt contain complicated language and grammar.

The results on the Text 2 document were around 50 percent which was a little less than what we expected, but fair enough given the drastically different nature of the text document as compared to Text 1.

VII. Conclusion

Through our proposed methodology of information extraction from raw, generic text documents, we were able to achieve fairly good results. As the generic nature and the complexity of language used in the Text document increases, we need to apply more advanced techniques in order to improve performance of our information extraction system. There are various methodologies which involve a machine learning approach to clean, summarize the text and extract information but those techniques require training of models on domain specific datasets which in turn compromises the generic nature of the solution. Hence for generic datasets, our proposed solution for information extraction gives fairly decent results.

References

[1] Ait-Mlouk and L. Jiang, "KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data," in IEEE Access, vol. 8, pp. 149220-149230, 2020, doi: 10.1109/ACCESS.2020.3016142.

[2] Gapanyuk, Y., Chernobrovkin, S., Leontiev, A., Latkin, I., Belyanova, M., , Morozenkov, O. (2018). A Hybrid Chatbot System Combining Question Answering and Knowledge-Base Approaches. AIST.

[3] Hien, Luong , Ly, Ly , Pham-Nguyen, Cuong , Le Dinh, Thang , Gia, Hong , Nam, Le. (2020). Towards Chatbot-based Interactive What- and How-Question Answering Systems: the Adobot Approach. 1-3. 10.1109/RIVF48685.2020.9140742.

[4] Shalaby, Walid , Arantes, Adriano , Gonzalez, Tere , Gupta, Chetan. (2019). Building Chatbots from large scale domain-specific knowledge bases: challenges and opportunities

[5] Ngai, Eric , Lee, Maggie , Luo, Mei , Chan, Patrick , Liang, Tenglu. (2021). An Intelligent Knowledge-based Chatbot for Customer Service. Electronic Commerce Research and Applications. 50. 101098. 10.1016/j.elerap.2021.101098.

[6] S. S. Japa and B. Rekabdar, "Memory Efficient Knowledge Base Question Answering with Chatbot Framework," 2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM), 2021, pp. 33-39, doi: 10.1109/BigMM52142.2021.00013.

[7] Georgios Patsoulis, Rafail Promikyridis, and Efthimios Tambouris. 2021. Integration of chatbots with Knowledge Graphs in eGovernment: The case of Getting a Passport. In ¡i¿25th Pan-Hellenic Conference on Informatics¡/i¿ (¡i¿PCI 2021¡/i¿). Association for Computing Machinery, New York, NY, USA, 425–429. https://doi.org/10.1145/3503823.3503901

[8] Lommatzsch, Andreas and Jonas Katins. "An Information Retrieval-based Approach for Building Intuitive Chatbots for Large Knowledge Bases." LWDA (2019).

[9] A. L. Krassmann, J. M. Flach, A. R. C. d. S. Grando, L. M. R. Tarouco and M. Bercht, "A Process for Extracting Knowledge Base for Chatbots from Text Corpora," 2019 IEEE Global Engineering Education Conference (EDUCON), 2019, pp. 322-329, doi: 10.1109/EDUCON.2019.8725064.

[10] Arsovski, Sasa , Osipyan, Hasmik , Muniru, Idris , Cheok, Adrian. (2019). Automatic knowledge extraction of any Chatbot from conversation. Expert Systems with Applications. 137. 10.1016/j.eswa.2019.07.014.

[11] Amato, Alessandra , Cozzolino, Giovanni , Ferraro, Antonino. (2021). Artificial Intelligent ChatBot for Food Related Question. 10.1007/978-3-030-61105-7-23.