Ontology Based Answer Extraction for Customer Support in Transportation Service

P.A.E.A Karunarathne[1], C.M.W. Arachchi[2], G. C. Perera[3], W.A.M.N Bandara[4] and G. U. Ganegoda[5]

Faculty of Information Technology, University of Moratuwa
Katubedda, Sri Lanka

[1]paeakarunarathne@gmail.com,[2]cmweerakoon@gmail.com,
[3]gethmaperera@gmail.com,[4]madushikanadeeshanfit@gmail.com,
[5]upekshag@uom.lk

Abstract: Customer support service providing facility using twitter platform is rapidly increasing. To enhance the service by reducing the time to provide the solution to the customer, is an important aspect in customer experience. To reduce the issues with the manual process, an automatic system is required with a knowledge base to extract answers for the raised questions by the customers. This research paper will discuss, a novel approach to design and the implementation an ontology-based answer extraction module for the transportation service. Using a question base and similarity matching before answer extraction will improve the accuracy of the extracted answer. The ontology was developed using Protégé and the answers are extracted using SPARQL query language.

Keywords: Answer Extraction, Customer support in Transportation Service, Ontology with Protégé

1 Introduction

With the rapid improvement of the digital world, the manual information processing, storing and retrieving became more tedious and labor-intensive. Therefore, developing automatic and accurate information storing and extracting systems became a popular research area among researches. Basically, Information extraction is automatic retrievals of certain types of information from natural language text [1]. Answer extraction is a type of information extraction that can be defined as retrieving exact information for a given question automatically without manual support. Mainly, web-based search engines and knowledge-based systems are implemented for answer extraction systems. Natural language processing techniques accurately understand the question and the answer is extracted from the web or the knowledge base.

Ontology is a formal and explicit specification of a shared conceptualization. The ontology-based answer extraction model is specifically designed to retrieve the exact answer for the questions with the use of ontologies. Ontology can be built as top-level ontology, task ontology, application ontology or domain ontology. Answer extraction models mostly use domain ontology due to its relativeness for the specific domain. Ontology-based answer extraction models are related to the natural language processing, text mining and knowledge representation [1]. Natural language processing and text mining are related to information extraction while ontology is related to the knowledge representation. Natural language processing techniques used to automatically understand the question and generate the answer. Text mining discovers previously unknown relations and information through data mining techniques. Ontology stores and processes knowledge. The goal of using ontology for answer extraction model is to obtain, describe and express related fields of knowledge [2] more accurately. Although the ontology is used to extract the correct answer, it is much difficult to find the exact answers due to the variation of the questions. The question varies in terms of synonymous, ambiguity, metaphors and many other ways. This problem should be addressed when constructing an ontology-based information extraction system. Thus, this research paper is discussed about the question-answer system related to the uber transport customer support service in twitter platform.

Twitter is one of the popular examples for a microblogging service provider. This platform is used to share opinions and questions on certain topics. Furthermore, due to its popularity, within a certain amount of time, an enormous amount of data can be extracted [3]. Therefore, these platforms are an essential source of information providers. More than 13% of microblogs on twitter are questions which involved needs of recommendation, opinions, request of information, etc. [4]. Organizations in different industries such as transportation, maintain a customer service providing a facility through twitter platform. Uber is one of the major transportation service providing companies and they maintain a twitter account to interact with their customers through @uber_support. The customers of the Uber provide reviews and raises questions they face using twitter. They use human resources to handle the customers. It is a challenge to create a system that provides the answer to the customer. In this study, for the knowledge base of the system, an ontology will be developed to extract answers for the raised questions by the customer, in the twitter platform. The tool Protégé is used for the development of the ontology and official Uber help page[1] is used as the source for information to the knowledge base.

The structure of the paper is as follows: section 2 discusses the related works about the ontology-based answer extraction systems. Section 3 explains the proposed approach for the ontology-based answer extraction systems. It is followed by the implementation of the system in section 4. Section 5 and 6 are followed by the discussion and the conclusion respectively.

2 Related Work

Present, the studies in the area of ontology-based answer extraction systems are the most popular topic among researchers. Because ontological understanding structures assume an imperative element inside the utility of question answering and data retrieval. This section described the

related works regarding the ontology-based answer extraction model.

S. Jayalakshmi and Dr. Anandhi Sheshasaayee [2] introduced Web And semantic knowledge Driven automatic question answering (WAD) approach for ontology-based query answering systems. The main goal of the WAD approach is to increase the accuracy of the document matching by analyzing and measuring the similarity between the words. This approach has used ontology as a common vocabulary to extends service for expanding the queries before submitting them to the search engine. This methodology has further improved the QA system, including relevant answer types based on the conditional probability and ontology structure [2].

Dr. P. Selvi Rajendran and Rufina Sharon [5] have completed research on the Dynamic Question Answering System based on Ontology in 2017. Seonyeong [6] used multiple strategies for both answer Natural Language (NL) questions and respond to keywords. The multiple information sources including knowledge base, raw text, auto-generated triples, and NL processing results. Morales [7] built up the Question Answering System, which gets to the data from Wikipedia data box and it accomplishes the high accuracy values. This work makes utilization of the Interface Agent to translate the methods for the buyer question and consequently bears remarkable choices.

In 2017, Kwong Seng Fong and Chih How Bong proposed a hybrid approach of automating the question and answer system. This is a combination of the knowledge-based approaches and text-based approaches. This approach requires two SPARQL to retrieve the answers from the ontology. They investigated and evaluated different language models and proposed a hybrid approach to improve the efficiency of answer retrieval [8].

Agnieszka Konys publish a research paper on knowledge systematization in the ontology-based information extraction domain. The implemented ontology allowed to capture and formalize ontology-based information extraction. This ontology is considered as an effective ontology for researches. This research paper provides a clear idea of ontology design and implementation [9].

Ankita Singh and Nidhi Tyagi proposed an approach to the efficient ontology-based question-answer system. This approach has a question processing module, query formulation module, and answer extraction module as the three main functional components. The architecture if this system is considered as simple and efficient architecture in terms of answer retrieval [10].

3 Proposed Approach

Fig. 1 describes the proposed approach for the answer extraction model for transportation service providers. First, the question is derived and sent to the preprocessing module to be preprocessed. The data is cleaned to remove duplicated, missing values and the unwanted phrases. It will be tokenized, stemmed and removed the stop words and retrieve the bag of words. Next, the words will be sent to the question classifier. Existing answer extraction ontologies will differ from this proposed approach because of the similarity matching and the question category base of this system. The similarity is matched with the words in the question base, in order to increase the accuracy of the retrieved answer. Wordnet is used for the similarity matching and the question base is specific to the domain of the answer ontology. Using the answer ontology, the answer will be extracted using the SPARQL query and it will be sent as the reply tweet.

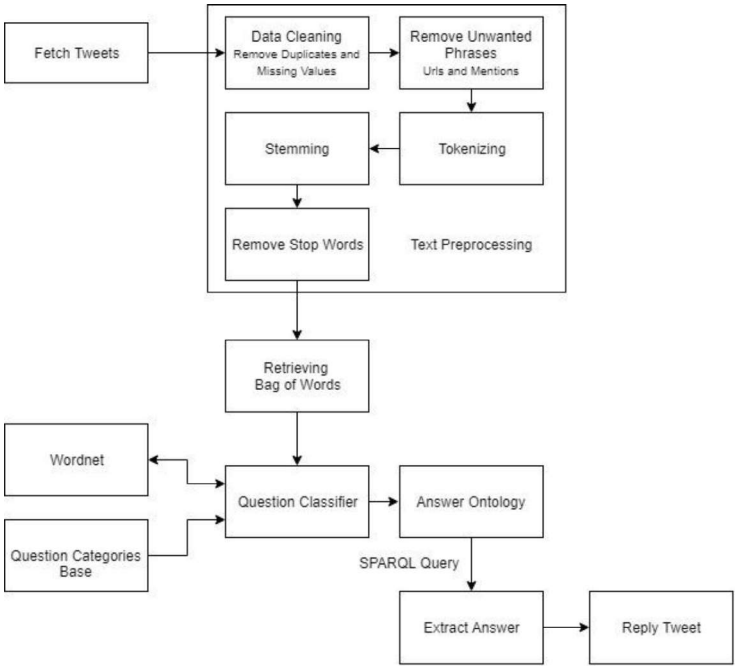

Fig. 21 Proposed Approach for the Ontology Based Answer Extraction System

3.1 Question Preprocessing

The questions are fetched in real-time from the Twitter API which includes the mention of the user profile. In this study, it is fetched from the users who have mentioned @uber_support in their tweets. The fetched tweets will be sent to the preprocessing unit. It will be cleaned to remove duplicates and the missing values. The unwanted phrases including URLs, other mentioned account names starting with @ are removed from the tweet. The library NumPy which is a fundamental package for scientific computing with python language is used for tidying the tweets.

Next, the tweet will be tokenized to a stream of tokens or words. For this research, the tokenization is done using the Term Frequency–Inverse Document Frequency (TFIDF). Necessary tokens will be derived using the TFIDF method by identifying the importance of each word in the collection in the transportation service providing domain.Moreover, the stream of tokens is then stemmed to derive meaningful words. The purpose of stemming the tokens are to remove suffixes, and to reduce number of tokens to accurately match each word which has the same meaning. Stop words are removed from the text to reduce the dimensionality in term of space and it is less important for the answer extraction process [11]. For both stemmer and for the removal of stop words, Natural Language ToolKit (NLTK) library is used. Finally in the preprocessing module, the bag of words is retrieved and sent to the question classifier.

3.2 Question Classifier

Retrieved bag of words is then sent to the Question classifier module which is the same bag of words used to identify the question. This classifier uses the WordNet database to match the similarities of the word by comparing a bag of words and identify a unique keyword set. A question category base consists of question categories that are predefined in Uber Help. Moreover, it contains common words related to user questions. This is defined by referring to the Uber Help site and the customer tweets. By referring to this WordNet database, the question is a map to the relevant property of the ontology. The algorithm is implemented with the help of cosine similarity to match similarities of words. The cosine similarity algorithm is based on Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF). For the similarity matching in this system, TF-IDF is chosen because it is recommended for search query relevance along with text similarity [12]. Cosine similarity considers the total amount of words to build the vectors according to the frequency (See equation 1). In the cosine similarity equation variable, A is the TF-IDF of the question and variable B is the TF-IDF of the keywords of each category [12].

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

3.3 WordNet Ontology

After the preprocessing stage, the question can be received as a set of words. The meaning of the question should be analyzed and mapped into question categories. These question categories are defined in a hierarchy, according to details of Uber help documentations. When mapping the question into the question hierarchy, considering synonyms of each word in the question, is much important. Because checking synonyms and find the similarities with a confidence level is an acceptable method rather than doing a string comparison and find similarities. Then the WordNet has applied as a solution for this question mapping problem [1].

WordNet is a database, which consists of many English words. It includes a set of synonyms (synsets), which are having cognitive synonyms and unambiguous concepts. These synsets are interrelated with each other based on their meaning and lexical relationships. WordNet called a lexical database and it's freely available to download. When considering these factors using WordNet is the best approach for similarity mapping. After extracting the synsets from the WordNet, similarities are checked. Then question mapping can be done considering the exact meaning of the question.

3.4 Answer Ontology

Ontology is defined as a formal and explicit specification of a shared conceptualization [1]. Ontology is domain knowledge used for semantic matching question answering system. The ontology for this answer extraction model is manually created by using Protégé. Protégé is a widely used ontology editor for modeling ontology. Web ontology language (OWL) used to define the ontology. In this process, the expected answers are identified as person, place, entity name model, passage retrieval and ranking, dates, currency or quantity model according to the preprocessing classifier at the question analysis model.

3.5 Performance Evaluation

When considering the performance evaluation in information extraction systems, precision and recall are the most popular performance evaluation metrics commonly used in the researches [1]. Precision provides the percentage of number of correctly identified answers of the total number of identified answers. Following equation 2 shows the precision.

$$\text{Precision} = \frac{|\text{Correctly identified answers}|}{|\text{Total answers}|}$$

Recall is the ratio between total number of correctly identified answers and number of questions. Following equation 3 shows the F1-measure metric.

$$\text{Recall} = \frac{|\text{Correctly identified answers}|}{|\text{Total number of questions}|}$$

Other than these two metrics, the special matric which is called F1-measure is also can be used to evaluate the

system. Following equation 4 shows the F1-measure metric.

$$\text{F1-measure} = \frac{2\,(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

4 Ontology Development

For the development of the ontology, Protégé tool was used. Because it enables to enter the class definitions, hierarchies, object properties and data properties in a customized way. The ontology is designed by referring to the Official Uber Help site. The top classes are identified as the question types defined in the site. WordNet is a freely available database. For the implementation, WordNet is used through python nltk library "corpus". The question base is also implemented by referring Uber to help web site. This consists of question categories that are predefined in the Official Uber Help site. Moreover, it contains common words related to user questions. This is defined by referring to the Uber Help site and the customer tweets. Table 1 depicts an example of scenarios of the question base. CategoryA and CategoryB are defined as per the website. CategoryA represents the main type of question or the general question category. And Category B represents the problem encountered or the question which provides the answer in the website.

Table 1. Example Scenarios Used in the Question Base

CategoryA	CategoryB	Key words	Related object properties
Account and Payment	Can't Request a ride	account, can't, request, ride	cantRequestARide
Account and Payment	Can't Sign in	account, can't, sign in, forgot, password	forgotPassword
Account and Payment	Can't Sign in	account, can't, sign in	cantSignIn

Top classes of the ontology were identified as the question types defined on the website. Fig. 2 illustrates the top-level hierarchy of the answer ontology. The 7 classes defined on the website are Account and Payment, Signing up, Guide, Accessibility, Other, Person and More. The related question domains are covered in the related classes. Table 2 depicts a set of object properties that are used in the ontology.

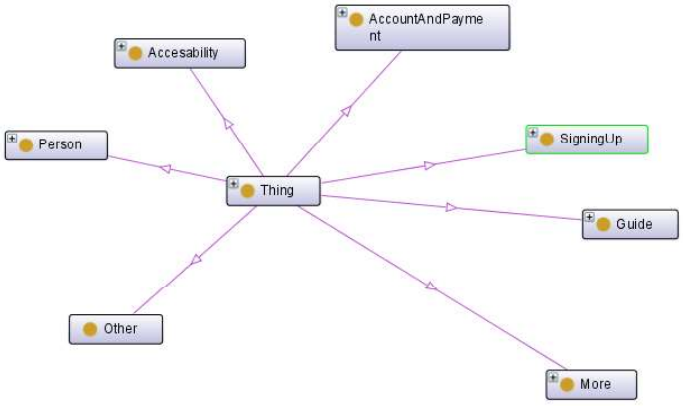

Fig 2. Top Level Hierarchy of the Answer Ontology

Table 2. Used Object Properties in the Ontology

Property	**Domain**	**Range**
cantSignIn	Customer	SignIn
cantRequestRide	Customer	RideRequest
declinePayment	Customer	PaymentDecline
useBackupCode	Customer	BackupCode
resetPassword	Customer	PasswordReset

After the implementation of the knowledge base, several scenarios were used to retrieve answers for raised questions. SPARQL query language was used to query and extract the answer. For the questions raised within the Uber transportation service domain, the answers that are included in the knowledge base are extracted using SPARQL. For the questions which do not include an answer in the knowledge base, are handled by sending the reply answer as "Please contact @Uber_Support via direct messages". Following figures represent the answers retrieved for sample scenarios. Fig. 3. Shows the answer to the question "How do I reset my password?" and Fig. 4. Shows the retrieved answer for the question raised "I can't request a ride".

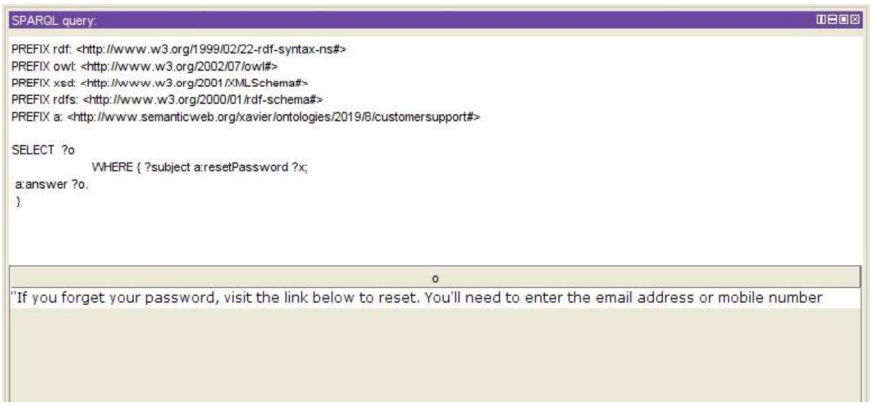

Fig 3. SPARQL query to retrieve answer for the question: "How do I reset my password?"

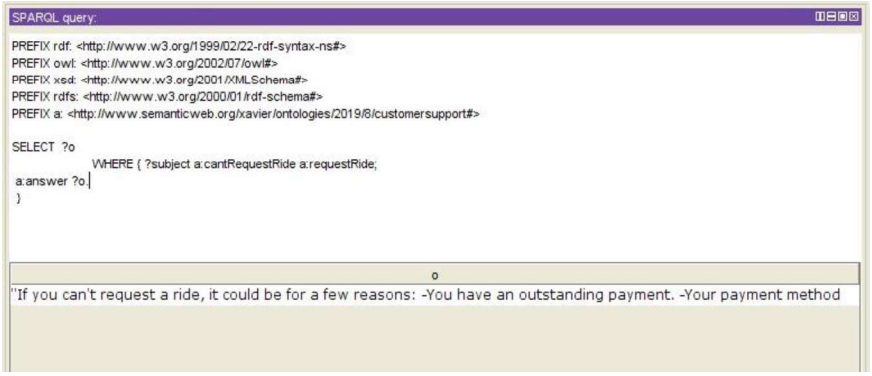

Fig 4. SPARQL query to retrieve answer for the question: "I can't request a ride"

5 Discussion

With the development of modern technology, information processing was also rapidly changed from manual to automatic. Developing an automatic system that is capable of the accurate store and extract information is a challenge in different domain areas. These types of answer extraction systems can be used as a customer support service providing tool in organizations. Twitter is one of the platforms which is used to provide customer support services. Different industry areas including transportation, security services, social services provide support to the customers when they raise questions, opinions, etc. Currently, organizations such as Uber uses human resources to handle their twitter accounts to handle the inquiries. A proper solution needs to be provided for the issues with human employees face in customer support services. Whereas, it is very important to improve the customer experience, reduce the time consuming to provide the solution to the customers.

This research paper mainly focused on the design and development of the knowledge base for the system by developing the ontology to extract the answers. Protégé is used for the development of the ontology. The official Uber help page is used as the source for information for the knowledge base. The text is extracted from the Twitter API and preprocess by cleaning, removing unwanted phrases, tokenizing, stemming and removing stop words. After retrieving the bag of words, it is sent to the question classifier where the similarity words will be matched with the use of the Wordnet. Similarity matching will improve the accuracy of the extracted answer in the evaluation process. Furthermore, similarity matching will be based upon the question categories base. It will be implemented specifically in the domain. Finally, the answer will be extracted using relevant SPARQL queries and will be sent as the reply tweet.

In the implementation phase, the class hierarchy, object properties, data properties were included with the reference of the Official Uber Help website. And the relevant question categories were defined as per the defined categories on the website. The retrieval of answers was checked using the SPARQL query language. And for the Uber transportation service domain, necessary answers were able to extract accurately. A question raised in the relevant domain; the area will be provided with the correct answer from the ontology. And the questions which are out of the domain will be sent a reply tweet to contact @uber_support via direct messages thought the Twitter platform.

Ontologies have been created in different research areas. But, in the customer, support service domain finding a developed ontology was a challenge. This research study covers the transportation customer support service and as the source for the knowledge base, leading transportation service providing organization, Uber official site was used. Since an already developed ontology did not exist in this domain area the required ontology needed to design and implement from the base. Therefore, it was a challenging task to design the ontology to satisfy all the possible questions which provide answers

according to the documentation. The individuals are manually added to the ontology. Another challenge was to choose the best similarity match algorithm for the ontology to provide higher accuracy in the similarity matching process.

This ontology can be used as the base ontology for transportation service providers who maintain their service via mobile or desktop application. And new knowledge can be added to the knowledge base by defining classes accordingly and including the necessary individuals to provide answers. There are existing answer retrieval ontologies developed. But for the transportation service domain, ontologies have not been developed. Moreover, to improve the accuracy, similarity matching will be done before the answer retrieval. And the question categories will be maintained separately in the question categories base to map the exact words used in the ontology.

6 Conclusions

In the transportation industry, providing customer support via twitter and other social platforms are rapidly increasing. the service via twitter is currently done by using human resources. To respond to a tweet that includes a customer problem takes a considerable amount of time. Therefore, it is important to enhance the customer experience. Using an automatic system to reduce the issues with the manual process is a challenge. This research paper includes a novel approach for answer extraction for the transportation service providers in the twitter platform. In this system, the question is fetched, preprocessed and retrieved the bag of words and sent to the question classifier for similarity matching. This is done with the use of the question categories base which is created specifying the transportation domain. This approach will improve the accuracy of the extracted answer. The question raised out of the domain is also handled by a general reply tweet. Although there are answer extraction ontologies in a different domain, there is no ontology that is created for customer support service in the transportation service area. This ontology can be used as the base for new ontologies that are related to the customer support service in the industry.

References

1. Daya C. Wimalasuriya , Ontology-Based Information Extraction, ,In: Journal of Information Science archive, Vol. 36 Issue 3, (2010)
2. Rohit Joshi, Rajkumar Tekchandani, "Comparative Analysis Of Twitter Data Using Supervised Classifiers", IEEE, 2016 International Conference on Inventive Computation Technologies (ICICT), (2016)
3. Miles Efron and Megan Winget, "Questions are Content: A Taxonomy of Questions in a Microblogging Environment", In: Journal Proceedings of the ASIST Annual Meeting, Vol. 47, (2010)
4. S. Jayalakshmi, Dr.Ananthi Sheshasaayee "Automated Question Answering System Using Ontology and Semantic Role", In: International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (2017)
5. Dr.P.Selvi Rajendran Rufina Sharon "Dynamic Question Answering System based on Ontology", In: IEEE International Conference on Soft Computing and its Engineering Applications (icSoftComp) (2017)
6. Seonyeong Park, Soonchoul Kwon, Byungsoo Kim, Sangdo Han, "Question Answering System using Multiple Information Source and Open Type Answer Merge", In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp 111–115 (2015)
7. Alvaro Morales, Varot Premtoon, Cordelia Avery, "Learning to Answer Questions from Wikipedia Infoboxes", In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 1930–1935 (2016)
8. Ankita Singh, Nidhi Tyagi, Ontology Based Question Answering System. In: International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 10, (2013)
9. Agnieszka Konys, Towards Knowledge Handling in Ontology-Based Information Extraction Systems, 22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Vo. 126, pp 2208-2218, (2018)
10. Kwong Seng Fong, Chih How Bong, A Hybrid Question Answering System based on Ontology and Topic Modeling, Vol. 9, No. 2-10 (2017)
11. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, Preprocessing Techniques for Text Mining - An Overview, In: International Journal of Computer Science & Communication Networks,Vol 5(1),7-16 (2015)
12. Gunawan, Dani & Sembiring, C & Budiman, Mohammad, The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. In: Journal of Physics: Conference Series. Vol. 978. (2018)