Search by Question (The SbQ System)

Mehmet Şerefoğlu– Koray Danışma– Ali Dorukhan Karakaya
Advisor: Dr. Engin Demir

Çankaya University, Department of Computer Engineering

Abstract

Finding correct answers to queries in search engines is still a problem as most of the search engines do search by keyword based interfaces. In this project, a question and answer type of intelligent search engine prototype is developed by exploring the data collection using interactive filtering to learn the actual intent of the user. Then, search by question system finds results of asked 5w1h question in the document collection and list them. Here, IBM Watson tool is used to process the data collection.

Introduction

The SbQ system is basis of semantic analysis that try to answer 5w1h questions. The main point here is natural language processing, NLP for short, is a method for system to analyze, understand, and make meaning from human language in a smart way. To produce a solution, NLP should be used efficiency and develop an useful algorithm to extract the possible answer(s) in returned data(s) from IBM Watson.

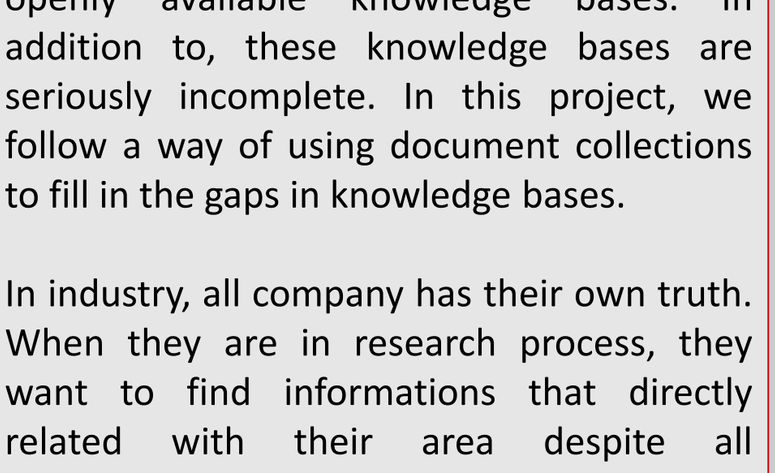

Figure 4 – User Interface

Industrial Issue

Over the past few years, massive quantity of world knowledge have been collected in openly available knowledge bases. In addition to, these knowledge bases are seriously incomplete. In this project, we follow a way of using document collections to fill in the gaps in knowledge bases.

In industry, all company has their own truth. When they are in research process, they want to find informations that directly related with their area despite all documents that contain their research key words.

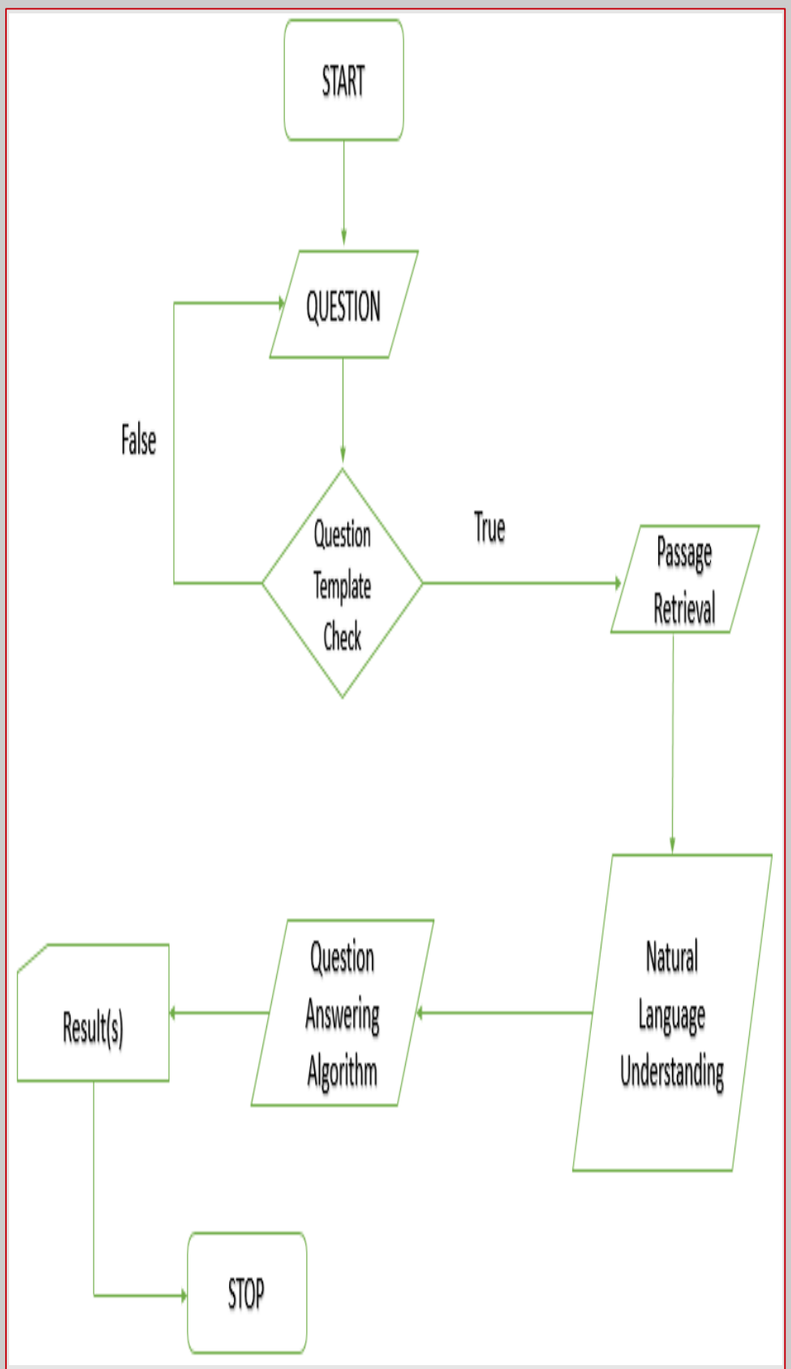

Figure 1 - Flowchart

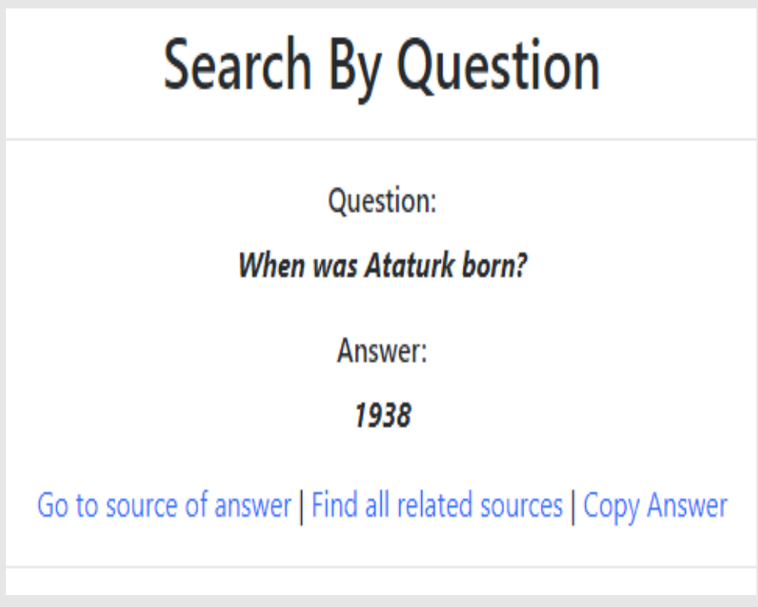

Figure 2 – Finished Product

Figure 3 – Training the NLU Tool

Solution

First step of solution process is getting the entered query from user interface. If it could be pass the question template check algorithm (then it is 5w1h question), IBM Watson Discovery returns related passage(s) that includes possible answers from data collection. The SbQ system extracts relevant information in to the corresponding entity structure. For instance, if a question term is “when”, the question type should be “date”. After specifying place(s) that the subject located in , we move on in each place from left of the subject and right of the subject to find related entity type at the same time. Thanks to IBM Watson Natural Language Understanding tool which is trained by The SbQ development team, system realizes and find the related entity type. However, it is not enough to return the found result(s). Because in same part there could be more than one data with related entity type. For example, for “date” entity type again, if user ask the born date of subject, we get the smallest date in the part and return it as the answer. In addition to alternative solutions, Discovery could find the verb in related passages. When it found the “born”, we return the first date following it as the answer to user.

Results & Conclusion

This project is seriously helpful for us to learn how to work with unstructured datas, analyze the machine learning algorithm and decide which one is appropriate to use in project, how to work in Natural Language Understanding process, how computers do semantic analysis and how we should train them.

In addition to, we learned how to work with web based program with using technologies that are php and nodejs.