Intelligent Question Answering System for HIPAA and Corporate Regulations

Vhinesh Ravi | vr8697rit.edu
Advisor : Dr. Carol Romanowski

Objective

Developing a closed domain cognitive question and answering system for HIPAA and Corporate regulations using Watson on IBM Bluemix Platform

Introduction

• The cognitive computing systems are a new class of systems that are designed to imitate the functionality of the human brain. The functioning of these systems is modeled after the functioning of the human brain.
• More then three fourth of all the data that is generated is unstructured lying in the form of text.
• Cognitive systems are designed to understand this unstructured data.
• The development of cognitive applications making use of these vast unstructured data is more than a necessity.
• Bluemix provides a platform for building the cognitive applications.

Document Preparation and Preprocessing

• The dataset comprises of 77 pdf text documents collected from the Arc of Monroe County website.
• The documents are broken into smaller documents paragraph wise.
• The documents are then converted into answer units where each answer unit has a unique id.
• The JSON is parsed for relevant fields and new answer units are formed with retaining the relevant fields.

Implementation

The Question and Answering System comprises of two parts Retrieve and Rank.

Retrieve:
• The retrieve component is created by building a Information retrieval engine through SOLR.
• The configuration file is created describing the schema of the data to be stored.
• Based on the search query relevant are documents retrieved using the term frequency – inverse document frequency score.

Rank:
• The Rank part is implemented by using the IBM Watson’s Learning to rank algorithm given as Ranker API service.
• The training data is created by forming 200 questions and mapping it with its associated answer id and defining its relevancy score.
• The Raker is trained using the training data.
• For a given input query, the ranker returns 10 most relevant answers.

Results

• The system is tested by creating 50 random test questions and asked against Ranker component as well as the Retrieve component.
• The answers returned by Ranker component were more relevant than answer returned by the Ranker component for 43 out of 50 test questions.
• Question — “Does arc reward employees?”

Conclusions and Future Work

• The System was successfully able to retrieve relevant answers for most of the test questions.
• The system retrieves 10 most relevant answers ordered in terms of their relevancy.
• The systems can be extended to other domains.
• The User feed back can be taken into account after each retrieval, So that the Ranker is able to do continuous learning dynamically.
• Making it into an API service, so that it can incorporated into other applications.

References