Interpretation of Natural Language using Data Mining,NLP and Machine Learning Techniques

CS4089 Project

End Semester Report

Anisha Kumari (B120535CS)
Drishya Praveen C P (B120086CS)
Indu Sree (B120668CS)
Guided By: Dr G. Gopakumar

24th February 2016

Abstract

Our project aims at building a prototype capable of answering questions posed in natural language, motivated by IBM Tool called WATSON. Watson has a knowledge base of structured and unstructured data. When posed with a question, it searches for the possible keywords in the query with what is present in its knowledge base. Normal keyword based search engines retrieves results which are optimistic yet they lack in their ability to interpret user question.

1 Introduction

A keyword based search engine presents to user enormous amount of data on being asked a question from which he or she cannot figure out the essential and most important information. Question answering system is a form of information retrieval system which aims to deliver the exact answer to the user question rather than whole document. To answer this user need semantic based reformulation techniques that can be used to retrieve the accurate answer from the enormous number of documents retrieved from the search engine. Watson is an artificial intelligent system that answers questions posed in natural language. It was developed by IBM specifically to answer questions in a quiz show called Jeopardy and it applies advanced natural language processing, information retrieval and machine learning to the field of question answering. It has millions of structured and unstructured data in its storage. It parses keywords while searching for related terms in a clue. It makes use Map-Reduce Algorithm.

It has Applications beyond Jeopardy too, such as :

- Watson can be used to fast-track life-saving research, create connections in cold cases all in a matter of seconds, not in weeks or months.
- Watson for Clinical Trial Matching : Uncover promising new approaches often unavailable outside of the clinical trial setting.
- Watson Curator : Optimize business information to increase user confidence in Watson responses.
- Watson for Oncology : Assistance for oncologists to make more informed treatment decisions.

2 Problem Statement

The problem is to develop a miniature prototype of IBM tool Watson using

Machine learning Techniques, Language Analysis Algorithms with the help of Apache Jena framework.

3 Literature Survey

3.1 WatsonPaths: Scenario-based Question Answering and Inference over Unstructured Information

WatsonPaths can answer scenario-based questions (The goal of scenario analysis is to identify information in the natural language narrative of the problem scenario that is potentially relevant to solving the problem), for example, medical questions that present a patient summary and ask for the most likely diagnosis or most appropriate treatment. It builds on the IBM Watson question answering system that takes natural language questions as input and produces precise answers along with accurate confidences as output. It breaks down the input scenario into individual pieces of information, asks relevant sub questions of Watson to conclude new information and represents these results in a graphical model i.e. assertion graph. Probabilistic inference is performed over the graph to conclude the answer and is repeated until a stopping condition is met. WatsonPaths takes a two-pronged approach to medical problem solving :

- First, by expanding the graph forward from the scenario in an attempt to make a diagnosis.
- Then linking high condence diagnosis with the hypothesis.

3.2 Ontology Based Information Retrieval System for Academic Library

The paper describes in detail the development of a search engine that interprets the meaning of users query instead of a keyword based search produces list of answers instead of specific answer. In this paper, an ontology based semantic information retrieval system is proposed utilizing Jena semantic web framework and Protege. A user enters an input query which is parsed by Stanford Parser followed by application of triplet extraction algorithm. Then SPARQL query is formed and it is fired on the knowledge base (ontology) which finds appropriate RDF Triples in knowledge base and retrieve the relevant information using Jena semantic web Framework.

3.3 Tools and Methods for Building Watson

This paper discusses the methodology followed by research team during the development of Watson. The paper describes in detail the software environment, software integration and testing protocol, the DeepQA computing environment built to support huge volumes of high-throughput experiments and the tools they built for system performance measurement and error analysis.

3.4 Natural language question answering: the view from here

It provides an overview of question answering as a research topic in terms of Applications, Users, Question types, Answer types, Evaluation, Presentation. There are also different methodologies for constructing an answer: through extraction, cutting and pasting snippets from the original documents containing the answer or via generation.This article discusses about 3 major activities :

- Information retrieval(IR)
- Information extraction(IE)
- Question Answering

IR techniques have been extended to return not just relevant documents, but relevant passages within documents. The IR community has, over the

years, developed an extremely thorough methodology for evaluation which resulted in development of recent question answering evaluation. IE may be viewed as a limited form of question answering in which the questions (templates) are static and the data from which the questions are to be answered are an arbitrarily large dynamic collection of texts.

It explains Generic Architecture for a Question Answering System and the steps involved in it :

- Question Analysis
- Document Collection Preprocessing
- Candidate Document Selection
- Candidate Document Analysis
- Answer Extraction
- Response Generation

4 Work Done in the previous semester

The literature survey is completed.

4.1 Design

Domain:
A short story : **"The Tiger King"**

Database:
The story in the form of unstructured text. Apart from the unstructured text database, there will be two tables maintaining all possible specific and generic questions.Under the generic questions table, we will be maintaining general questions regarding the details of the story and its book like - name, author, name of the publisher etc and its answers. In Specific questions table, we will be writing all possible questions and summary corresponding to each paragraph of the book.

Ontology:
Ontological database for storage of book information using Protege tool is created and this database is used to retrieve related answer from the domain specific ontology.
User interface is created in which user enters input question in natural language.
When a question is posed, it is first searched in both Generic and Specific Questions table and if a match is found,then a corresponding answer is returned to the user else proceed with the following steps.

SYNTAX ANALYSIS

Tokenization

Input: Question

Process:
Question is subdivided into tokens using tokeniser which includes lemmatisation using NLTK library by making use of Snowball stemmer. It is written in Python language.

Output: Tokens

POS Tagging

Input: Tokens

Process:
Each token is associated with its part-of-speech tags using NLTK. Some of the POS tags are NN - Noun Singular VB - Verb Base Form NNS - Noun Plural DT -Determiner

Output: Tokens with their corresponding POS tags

Chunking

Input: Tokens with POS tags,Chunk Grammar

Process:
Chunk Parser is formed using Chunk Grammar which segments and labels multi-token sequences. Chunk Parser is used to construct tree structure of question.

Output: Parse Tree

SEMANTIC ANALYSIS

WordNet is a lexical database for the English language. WordNet can help a question answering system to identify synonyms. The synonym information can be used to help match a question with an appropriate rule.

RDF Triple Extraction

Input: Parse Tree

Process :
Triplet Extraction Algorithm is used to extract Subject, Predicate, Object from the tree structure.

1. A sentence (S) is represented by the parser as a tree having three children: a noun phrase (NP), a verbal phrase (VP) and the full stop (.). The root of the tree will be S. Firstly we intend to find the subject of the sentence. In order to find it, we are going to search in the NP subtree. The subject will be found by performing breadth first search and selecting the first descendent of NP that is a noun.

2. Secondly, for determining the predicate of the sentence, a search will be performed in the VP subtree. The deepest verb descendent of the verb phrase will give the second element of the triplet.

3. Thirdly, we look for objects. These can be found in three different subtrees, all siblings of the VP subtree containing the predicate. The subtrees are: PP (prepositional phrase), NP and ADJP (adjective phrase). In NP and PP we search for the first noun, while in ADJP we find the first adjective.

Output: Triple

SPARQL Query Generation

Input: Triple

Process:
Before generating SPARQL query, WordNet or local dictionary can be used to identify synonyms for predicates. If matching predicate is not found in RDF database, Query is generated using Apache Jena framework.

Output: SPARQL query

Information Extraction

Input: SPARQL Query, Ontology

Process:
Jena provides SPARQL API to handle both SPARQL query and their update. Jena API enables the SPARQL query for mapping with RDF database then it is fired on RDF database and retrieves the relevant information performing semantic search into database.

Output: Answer

5 Work Done in the current semester

- We took a short story from NCERT-XII "The Tiger King". All possible questions from each paragraph with their answers were made and a database was populated. After this a User Interface was created (using HTML, PHP, CSS) where a user can get answers to the direct questions if it is already present in the database.
- We wrote an algorithm to find the mood of a passage and imple-

mented it using Bag of Words on random passages. Sample answers are like Mood is Positive / Negative / Neutral. For implementing this algorithm, we created separate databases of large number of positive and negative words. When a passage is given as input, the algorithm calculates the number of positive and negative words and displays the mood accordingly.

- We have completed Syntax Analysis phase using Stanford Parser and NLTK. We performed operations like Tokenization, Chunking, POS Tagging on the input questions for getting the parse tree.

- We are currently working on generating RDF Triples using Triplet extraction Algorithm in the Semantic Analysis phase. We have already implemented a part of Triplet Extraction Algorithm which finds Subject, Object, Predicate of the input questions.

- For finding Tone of a passage, we have created a database. It contains many passages with their tone which will act as a Training set.

6 Future Work and Conclusions

The future work includes getting familiarised with Protege and Apache Jena framework to develop ontology for completing the Semantic Analysis Phase. We will be working to write an algorithm to find tone of any random passage using Naive Bayes Method. We will also be implementing programs for finding Moral and Scope of any given passage.

References

[1] Adam Lally, Sugato Bachi, Michael A. Barborak, David W. Buchanan, Jennifer Chu-Carroll, David A. Ferrucci*, Michael R. Glass, Aditya Kalyanpur, Erik T. Mueller, J. William Murdock, Siddharth Patwardhan, John M. Prager, Christopher A. Welty. *WatsonPaths: Scenario-based Question Answering and Inference over Unstructured Information*, RC25489 (WAT1409-048) September 17, 2014.

[2] Amol N. Jamgade and Shivkumar J. Karale, *Ontology Based Information Retrieval System for Academic Library* ,IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICIIECS) 2015.

[3] Eric Brown, Eddie Epstein, J. William Murdock, Tong-Haing Fin, *Tools and Methods for Building Watson* , RC25356 (WAT1302-021) February 15, 2013.

[4] L.Hirschman, R.Gaizauskas, *Natural language question answering: the view from here* , 2001 Cambridge University Press.

[5] Haiqing Hu, Peilin Jiang, Fuji Ren and Shingo Kuroiwa, *Web-based Question Answering System for Restricted Domain Based of Integrating Method Using Semantic Information* .