

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

CS4089 Project

Final Presentation

Anisha Kumari B120535CS

Drishya Praveen C P B120086CS

Indu Sree B120668CS

Guided By: Dr G Gopakumar

May 10, 2016

Outline

Introduction

Problem Statement

Literature Survey

Work Done

References

Introduction

- ▶ Extracting Keywords:
 - ▶ A keyword based search engine - results in enormous amount of data available to the user, from which user cannot figure out the essential and most important information. [Limitation]
- ▶ Question Answering : [semantic search]
 - ▶ IR system - exact answer to the user question.
 - ▶ Semantic based Reformulation Techniques - accurate answer from the enormous data retrieved from the search engine.
- ▶ IBM Watson:
 - ▶ Cognitive machine - Thinking machine like Humans.
 - ▶ Machine learning model - learns over time with reasoning model at base.

Problem Statement

- ▶ The problem is to develop a Question Answering system inspired from IBM tool Watson using Machine learning Techniques and Language Analysis Algorithms with the help of Apache Jena framework.

Literature Survey

- ▶ Ontology Based Information Retrieval System for Academic Library.
 - ▶ Development of a Search Engine :
 - ▶ Interprets meaning of query instead of a keyword based search
 - ▶ Specific answer instead of List of answers.
 - ▶ Ontology based semantic Information Retrieval System : Jena semantic web framework and Protege.
 - ▶ Triplet Extraction Algorithm - Parse Tree
 - ▶ SPARQL query formed - fired on the knowledge base (ontology), finding appropriate RDF triples and retrieve relevant information using Jena semantic web Framework.

- ▶ Natural language question answering: the view from here.
 - ▶ 3 major activities : Information retrieval(IR), Information extraction(IE) and Question Answering.
 - ▶ Generic Architecture for a QA System and steps involved - Question Analysis, Answer Extraction, Response Generation, etc.

Work Done

- ▶ Domain : Short Story "The Tiger King"
- ▶ Database : The book is in the form of unstructured text. Apart from the unstructured text database, there is a table maintaining all possible specific and generic questions.
- ▶ Ontology : Ontological model for storing book information is built using Protege tool which is later used to retrieve answer from the ontology.

► Syntax Analysis

► Tokenisation

- Input : Question
- Process : Question is subdivided into tokens using tokeniser in NLTK library and further stemmed with the Snowball stemmer.
- 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, VB, NNS, DT.
- 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 the question.
- Output : Parse Tree

- ▶ Semantic Analysis
 - ▶ RDF Triplet Extraction
 - ▶ Input : Parse Tree
 - ▶ Process : Triplet Extraction Algorithm is used to extract Subject, Predicate, Object from the tree structure.
 - ▶ Output : Triplet
 - ▶ SPARQL Query Generation
 - ▶ Input : Triple
 - ▶ Process : Query is generated using Apache Jena framework.
 - ▶ Output : SPARQL query
 - ▶ Information Extraction
 - ▶ Input : SPARQL Query, Ontology
 - ▶ Process : Jena provides SPARQL API to handle SPARQL query which is then fired on RDF database and retrieves the relevant information performing semantic search.
 - ▶ Output : Answer

- ▶ Made a User Interface for entering the queries.
- ▶ Created a database for answering direct questions.
- ▶ Created ontological database for the short story using Protege tool.
- ▶ Completed Syntax Analysis using Stanford Parser and NLTK to get Parse Tree.
- ▶ Generated RDF Triples using Triplet Extraction Algorithm in the Semantic Analysis phase.
- ▶ Using the Triplets, formed a SPARQL query using Protege tool.
- ▶ Found the Mood of a passage(Bag of Words) and Tone of a passage(Naive Bayes).

References I

- [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.