Ontology Based Question Answering using Semantic Similarity Matching

You

January 20, 2016

Abstract

Question Answering can be improved by focusing on three areas like the ontology enhanced processing and augmentation, content manipulation approaches, the query and the answer. Ontology enhanced processing could enhance answers to include identified objects satisfying a query. Natural language user questions and information sources with a common ontology are required for ontology-based QA systems. A Question Answering System returns answer to a user question in succinct form. In order to provide a precise answer, the system must know what exactly a user wants. The prior knowledge of the expected answer type helps the Question Answering System to extract correct and precise. Question Answering is one of the major issues in e-learning research on how to provide more interactive activities around the learners and instructors. Every answer to the questions must be relevant to the users query in that context.

The input is given to the tree-tagger parser to identify the syntactical information. This syntactical information gives us the lexical constraints like Noun NN, Verb VV and other terms. The noun and verb keywords are analyzed with the semantic meaning using WordNet and semantic similarity measures. This paper proposes a method for QA system by providing different patterns for the same questions.

1 Introduction-Research Overview

Semantic Web and Ontology has been chosen as research area as this is evolving drastically and will be the future technology.

All the ongoing projects are developed based on Web 2.0. Down the line or in next few years all the projects are developed / converted to Web 3.0. Most important feature in Web 3.0 is Semantic Web and Personalization. It also enables the use of autonomous agent to perform some task for the user. This makes search engines to search based on user needs rather than the keywords.

The current Research and Development Focus of the United States - Department of Defense's Advanced Distributed Learning (ADL) initiative describes these autonomous agents as the personal assistants to the Next Generation Learner in the Next Generation Learner Environment. The Personal Assistant for Learning (PAL) is a long-term focus of ADL's R and D endeavors over the next few years.

The goal of ADL's research is to create a capability that anticipates learner needs, seamlessly integrates yet-to-be available information, and provides everywhere access to effective, personalized learning content and/or job performance aids that can be accessed from multiple non-invasive devices and platforms.

I feel this is the right time to explore and work on this technology as the importance is gaining more on Web 3.0. The aim of this research work is "to search the answer for the given question in ontology domain using JSON-LD with semantic similarity approach".

2 Related Works

2.1 E-LEARNING

In an e-Learning environment there is a high risk that two authors express the same topic in different ways. The problem could be solved using domain (content) ontologies in which mappings from domain vocabulary(s) in the commonly-agree terms are defined extensionally

Learning material could be presented in the various learning contexts or in the various presentation contexts. The context description enables context relevant searching for learning material according to the preferences of the user. In order to achieve shared-understanding about meaning of the context vocabulary a context-ontology is used. Because e-Learning is often a self-paced environment, training needs to be broken down into small bits of information that can be tailored to meet individual skills gaps and delivered as needed. These chunks of knowledge should be connected in order to create the whole course. The structure isn't a static one, because it depends on user type, users' knowledge level, users' preferences and prerequisite materials.

2.2 ONTOLOGY And E-LEARNING

Ontology provides a common vocabulary, and an explication of what has been often left implicit. Indeed, ontologies are a means of specifying the concepts and their relationships in a particular domain of interest. Web Ontology languages, like OWL, are specially designed to facilitate the sharing of knowledge between actors in a distributed environment.

From the modeling point of view, ontology languages are not only able to integrate Learning Object Model and Dublin Core metadata, but also allow for the extension of the description of the learning objects with non-standard metadata, thus giving users and groups of users more flexibility when sharing resources. Ontologies can be used in e-learning as a formal means to describe the organization of universities and courses and to define services. An e-learning ontology should include descriptions of educational organizations (course providers), courses and people involved in the teaching and learning process. Ontology is one of the essential methodologies for representing domain-specific concepts.

2.3 NATURAL LANGUAGE PROCESSING

NATURAL LANGUAGE PROCESSING

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at

one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

'Naturally occurring texts' can be of any language, mode, genre, etc. The texts can be oral or written. The only requirement is that they be in a language used by humans to communicate to one another. Also, the text being analyzed should not be specifically constructed for the purpose of the analysis, but rather that the text is gathered from actual usage. The notion of 'levels of linguistic analyses refer to the fact that there are multiple types of language processing known to be at work when humans produce or comprehend language. It is thought that humans normally utilize all of these levels since each level conveys different types of meaning. But various NLP systems utilize different levels, or combinations of levels of linguistic analysis, and this is seen in the differences amongst various NLP applications. This also leads to much confusion on the part of non-specialists as to what NLP really is, because a system that uses any subset of these levels of analysis can be said to be an NLP-based system. The difference between them, therefore, may actually be whether the system uses 'weak' NLP or 'strong' NLP.

The goal of NLP as stated above is "to accomplish human-like language processing". The choice of the word 'processing' is very deliberate, and should not be replaced with 'understanding'. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU system would be able to paraphrase an input text, translate the text into another language, answer questions about the contents of the text and draw inferences from the text. Humans have been shown to use all levels of language like Phonology, Morphology, Lexical, Syntactic, Semantic, Discourse and Pragmatic to gain understanding, the more capable an NLP system is, the more levels of language it will utilize. Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches have coexisted since the early days of this field.

2.4 QUESTION ANSWERING

Question answering systems are designed to find answers to open domain questions in a large collection of documents. Using Google we find documents containing the question itself, no matter whether or not the answer is actually provided. Current information access is query driven. Question Answering proposes an answer driven approach to information access.

Level One questions cause students to recall information. This level of question causes students to input the data into short-term memory, but if they don't use it in some meaningful way, they may soon forget.

Level Two questions enable students to process information. They expect students to make sense of information they have gathered and retrieved from long-and short-term memory.

Level Three questions require students to go beyond the concepts or principles they have learned and to use these in novel or hypothetical situations. Many years ago, an educator named Benjamin Bloom developed a classification system we now refer to as Bloom's Taxonomy to assist teachers in recognizing their various levels of question-asking. Taxonomy is an orderly classification of

items according to a systematic relationship (low to high, small to big, simple to complex). Blooms framed the levels as Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation.

A question answering system may be open-domain or closed domain. Open-domain question answering deals with questions about everything, and it relies on general ontologies. On the other hand, Closed-domain ones deal with questions under a specific domain (for example, tourism or hardware gadgets), and can be an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in ontologies.

2.5 TREE-TAGGER And LINK GRAMMAR

There are components used like Link Grammar or Tree-Tagger, WordNet, Ontology, JSON–LD, etc.

2.5.1. TREE TAGGER

The Tree-Tagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The Tree-Tagger has been successfully used to tag German, English, French, Italian, Spanish, Russian, Greek, Chinese, Latin, etc. is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its contexti.e., relationship with adjacent and related words in a phrase, sentence, or paragraph.

A simplified form of this is commonly taught school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Tree-Tagger Parsers are used to extract the noun phrase, verb phrase for the current question answering system. The following table 2-1, 2-2, A2-3 gives the output of the parser.

2.5.2. LINK GRAMMAR Link grammar which developed by Carnegie Mellon University is a graphical grammar analyzing tool. Link grammar is a context-free formula to describe natural language. This system can produce all grammar linkage from English sentence which users input and determine the sentence correctness thought the linking result.

Figure 2-1 will show what Link Grammar is. First, input an English sentence into this system. Each word has some curves and each curve has one label on it. The curve and label is a Link which expresses a kind of linkage. After analyze and parse the sentence though Link Grammar, we may get a lot of Linkages. This information can help to realize the syntax structure of the question.

2.6 WORDNET

WordNet originated from Cognitive Science Laboratory in Princeton University. It is a vocabulary reference system designed by researchers who inspired by psychology theory.

WordNet is also referred as lexical database which is available online and provides a large repository of English lexical items. There is a multilingual WordNet for European languages which are structured in the same way as the English language WordNet.

Table 1: EXPERIMENTS RESULTS

Data set	Questions/Answers	Semantic Similarity
Q1 A1	What is the biggest city in USA New York is the biggest city in USA	1.0
Q1 A2	Give the biggest city in USA New York is the biggest city in USA	0.87
Q1 A3	Give the name of biggest city in USA New York is the biggest city in USA	0.76

WordNet processed the first level classification according to part of speech (POS) tag to establish the connections between four types of part of speech – noun, verb, adjective and adverb. Driven by different word meanings and expressions, it forms several Synset. Each Synset symbolizes one vocabulary and takes down other words and expression with the same meaning. Synset is the smallest unit in WordNet.

The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents.

For example, the words night, nighttime and dark constitute a single synset that has the following gloss: the time after sunset and before sunrise while it is dark outside.

Synsets are connected to one another through the explicit semantic relations. Some of these relations (hypernym, hyponym for nouns and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies.

For example, tree is a kind of plant, tree is a hyponym of plant and plant is a hypernym of tree. Analogously, trunk is a part of a tree and we have that trunk as a meronym of tree and tree is a holonym of trunk. For one word and one type of POS, if there is more than one sense, WordNet organizes them in the order of the most frequently used to the least frequently used (Semcor).

2.7 How to Write Mathematics

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

2.8 How to Make Sections and Subsections

Use section and subsection commands to organize your document. LaTeX handles all the formatting and numbering automatically. Use ref and label commands for cross-references.

2.9 How to Make Lists

You can make lists with automatic numbering ...

1. Like this,
2. and like this.

. . . or bullet points . . .

- Like this,
- and like this.

. . . or with words and descriptions . . .

Word Definition

Concept Explanation

Idea Text