A Survey on Natural Language Query Processing and Semantic Image Search

Rashmi Patel
Department of Computer Engineering, Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat, India

***Abstract* - The Field of Natural Language Processing (NLP) is becoming one of the active areas in Human-machine interaction which has seen a study shifted in both methodology and research direction in the past few years. The main purpose of Natural Language Query processing is to understand how computer interprets the query and perform appropriate action for them. Natural Language is convenient method of accessing data, especially for casual users who do not understand the technical way of searching. This paper also focuses on the semantic concepts to improve the relevancy of Image search.**

I. INTRODUCTION

World Wide Web (WWW) has enabled users to access information from all over the world. Information available on the web considered as vast, so the burden on the user to search, filter and select the desired information increases drastically. One possible solution is to enhance the infrastructure of www by adding the semantic concept to the search. This is the motivation for the semantic web which inherits the decentralized architecture of traditional keyword based search. Semantic web extends the World Wide Web with semantics, which adds meanings to the terms appear in web search query. To understand the Natural Language processing and Semantic fundamental for web search are current trend.

The Contribution of This paper:

A. Introduction to NLP
B. Introduction to Semantic Search
C. Semantic Measures to calculate the Similarity between Query terms
D. Sematic Image Search

II. NATURAL LANGUAGE PROCESSING

Natural Language Processing is a field of Artificial Intelligence which concerned with the interaction between computer and Human Language [1]. NLP involves complete understanding human utterance, so it can give response to them. The goal of NLP is to design and build software that analyze, understand and generate language that human use naturally, so that eventually you will be able to address your computer as though you were addressing another person[2]. NLP involves the study of Information Retrieval, Information Extraction, Machine Translation and Language Analysis. Fig.1 shows the query processed in Natural Language Environment to provide interaction between user query inputs which can be machine processable.

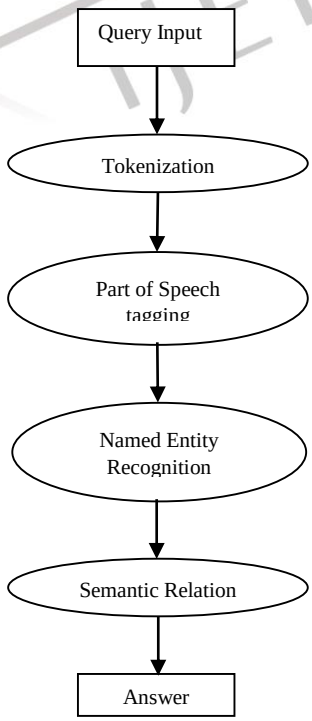

Fig.1 Query Processed in Natural Language Environment

In Natural Language Processing Environment, first the query is split into number of words called tokens which is performed using tokenizer. Then those tokens are tagged with part-of-speech tags, which will identify the tag like noun, pronouns, adjective etc. Then next step is to identify the entities using named entity recognition. Last step is to establish the semantic relation between the entities.

III. SEMANTIC SEARCH

Semantic is defined as the meaning or interpretation of word or sentence. Semantic search involves the study and implementation of semantics in the search technology in order to find out the real intent of the searcher behind the search query and presenting the set of results that closely relates to what the user is searching. It identifies a proper relationship between the terms used in search query before presenting the final search results[3].

Various Factors to understand Semantic search:

- Current trend
- Location of search
- Intend of search
- Variation of words
- Synonyms
- Generalized and specialized search
- Concept matching
- Change of meaning based on group of words

IV. SEMANTIC MEASURES TO CALCULATE SEMANTIC SIMILARITY

Semantic measures are widely used today to compare semantic entities such as units of concepts, language or even semantically characterized instances, according to information supporting their meaning. Semantic measures are used to measure the semantic similarities between words and sentences.

Semantic Similarity

Semantic similarity is a metric which define over a set of terms[5]. This is typically established by analyzing a set of list or document of terms and assigning a metric based on the concept they represent or express. To be semantically similar, two terms do not have to be synonyms one but they should represent same meaning or relate with each other. Semantic similarities between entities change over time with changing the domains. Semantic similarity can be calculated between words and sentences using semantic measures.

Semantic Similarity between words

Semantic similarity between words can be measured using two semantic measures such as Page counts and Snippets. Page counts and Snippets are web search based measures for finding the similarity between words.

Page Counts based Measures:

Page counts calculate number of pages that contains the keyword and also performs co-occurrence of keyword in a page. It finds co-occurrence between keywords in two ways: individual co-occurrence of two keywords and both together. The co-occurrence of words using Page counts is calculated using following measures such as WebJaccard, WebOverlap, WebDice and WebPMI.

Snippet based measures:

Snippet extracted from search engine for the conjunctive query which represents the local context in which two words co-occur on the web. The numerous semantic relations are identified from snippet. Numerous lexical syntactic patterns are identified and the frequency of the pattern is calculated. The lexical patterns that are conveying the similar semantic relations are clustered together to efficiently represent the semantic relation between words[4].

This similarity measures establish the relationship between words that two words can be Synonyms, Antonyms, Polysemy and Homophony of each other.

Synonyms: Similar meaning between words.

Antonyms: Opposite meaning between words.

Polysemy: Word which has two or more related meaning.

Homophony: When two meanings are entirely unrelated.

Semantic Similarity between sentences

Sentence is considered as bag of words to calculate the similarity between them. The relation between two sentences mainly falls into two categories: The semantic equivalence and the topic relevance[10].

The semantic equivalence is further divided into two types:

Paraphrase: Two sentences are paraphrase of each other; even though the meaning is different you will understand the same meaning from them.

Entailment: Two sentences don't mean exactly the same thing.

1) Mutual Entailment: Each sentence must be true for other to be true.
2) Asymmetrical Entailment: Only one sentence must be true for other to be true.

The topic relevance includes three levels: on the specific topic, general topic and unrelated.

V. SEMANTIC IMAGE SEARCH

In Web Image Search there are many pieces of information which can be used for ranking images. It is very difficult to identify the most relevant Image according to the user's Query. Traditional or Keyword based Image search doesn't provide the relevancy while retrieving the image. Therefore, Semantic concept applied on Query input which retrieves more relevant Images. This Scenario of retrieving those relevant images is called Semantic Image Search.

VI. CONCLUSION

This paper is a survey on Natural Language processing and how query is processed in Natural Language Environment. It also covers the understanding of Semantic concepts used in search engines to improve web search result as well as Image search.

VII. REFERENCES

[1] https://en.wikipedia.org/wiki/Natural_language_processing
[2] Gauri Rao, Chanchal Agarwal, Snehal Chaudhry, Nikita Kulkarni, Dr. S.H. Patil, "Natural Language Query Processing Using Semantic Grammar'' International Journal on Computer Science and Engineering
[3] https://en.wikipedia.org/wiki/Semantic_search
[4] Lakshay Sahni, Anubhav Seghal, Shaivi Kochar, Faiyaz Ahmad, Tanvir Ahmad . "A Novel Approach to Find Semantic Similarity Measure between Words'' International Symposium on Computational and Business Intelligence , IEEE 2014
[5] S. Lavanya and S.S.Arya "An Approach for Measuring Semantic Similarity between Words Using SVM and LS-SVM" International Conference on Computer Communication and Informatics, IEEE 2012
[6] Rafael Ferreira, Rafael Dueire Lins, Fred Freitas, Bruno Avila, Steven J. Simske, Marcelo Riss . "A New Sentence Similarity Method based on Three-Layer Sentence Representation'' International Joint Conference on Web Intelligence (WI) and Intelligent Agent Technologies, IEEE/WIC/ACM 2014
[7] Debajyoti Mukhopadhyay, Manoj Sharma, Gajanan Joshi, Trupti Pagare, Adarsha Palwe "Experience of Developing a Meta-Semantic Search Engine" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies , IEEE 2013
[8] Bernard J Jansen, Danielle L.Booth , Danielle L. Booth, Amanda Spink " Determining the User Intent of Web Search Engine Queries" Information Processing and Management, Elsevier 2008
[9] Eugene Agichtein, Steve Lawrence, Luis Gravano "Learning Search Engine Specific Query Transformations for Question Answering" ,ACM 2010
[10] Xiao-Ying Liu, Chuan-lun Ren, "Similarity Measures Based on Sentence Semantic Structure for Recognizing Paraphrase and Entailment" International conference on Machine Learning and Cybernetics, IEEE 2013
[11] http://www.searchenginejournal.com/seo-101-semantic-search-care/119760/