International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-5, August 2015
ISSN: 2395-3470
www.ijseas.com

An Approach for Representing Knowledge in Natural Language Processing

gnknithin@gmail.com
Gollapalli Nithin Kumar
M.Tech, CSE Dept.,
JNTUA College of Engineering, Anantapuramu, A.P, India

Abstract

Natural Language Processing is the process of providing interaction between computer and human in linguistic concerns. Knowledge Representation is designing computer systems to perform tasks that would normally require human intelligence. Both of these belong to the fields of Computer Science, Artificial Intelligence which can bring forth the knowledge and make it explicit. This can lead to develop Expert Systems with human like intelligence like Question-Answering, Automated Reasoning, Machine Translation systems that can work on natural language. Expert systems are used for decision making ability based on stored facts. Expert systems are a kind of knowledge based systems that are dependent on inference rules. Inference rules are modelled with either First Order Logic or Prepositional Logic, which can perform forward and backward chaining for deriving reason of any given query based on stored or existential facts. These existential facts are real facts or knowledge for the system and this is an approach for acquiring and representing knowledge from real-time feed.

Keywords—Artificial Intelligence, Natural Language Processing, Information Extraction, Knowledge Representation.

1. Introduction

With Watson [1], the world has seen how a computer system is able to compete at human in real-time. The advancement of a computer system to such an extent is made possible with the IBM research project DeepQA [2]. It is a software architecture that performs deep content and evidence based analysis for a question posed in natural language. To perform such a task, it needs most advanced natural language processing, semantic analysis, information retrieval, automated reasoning through machine learning. For such kind of content analysis it uses a software framework called Unstructured Information Management Architecture also called as Apache UIMA [3]. It is the only industrial software framework used to perform content analysis over large volumes of structured and unstructured data to find required knowledge. This works on different levels of components and its interfaces through an analytical pipeline, finding the relevant design patterns, organizing them in memory and convert them into structured data, so that the answer can be generated by building hypothesis based on resources.

Text Mining and Information Extraction are the major tasks performed by this architecture. Text Mining is a process of performing search, index, labelling, etc. Information Extraction is a task of extracting structural, semi-structural, unstructured, machine readable data from the documents. Most of the cases, these documents are presented in natural languages that are human readable context. Extracting information from the documents that are in natural language is challenging. Such task can be accomplished using Natural Language Processing.

The greatest challenge in the history of computer machinery and intelligence is Turing Test. It is proposed by Alan Turing in which a machine is said to be intelligent when it is able to answer the questions posed in natural language with limited available resources without aid. This test has many controversies and the research is always performed to produce, intelligent machine with specific kinds of purpose and limitations. The intelligence that is aimed to accomplish for machines or computers is called as Artificial Intelligence. Artificial Intelligence is defined as the designing of intelligent agents that take rational decision just like humans. It includes learning, planning, knowledge, reasoning, communication as human perspective, but according to machine approaches it deals with problem solving methods, machine learning, knowledge representation, decision making, etc.

UIMA is one good industrial standard that exists for logistic analysis system software. There are other frameworks like General Architecture for Text Engineering (GATE) and Natural Language Processing Toolkit (NLTK) at present. GATE is developed to perform wide range of natural language tasks, information extraction with the help of the Java programming language which is widely used by the scientific community as well as in academics for

better approached in their tasks. NLTK is a suite of libraries and programs that contain statistically based approaches using Python programming language. With the help of NLTK the process of text mining, information retrieval, etc., can be achieved.

Computers are essential tools for humans to deal with existing information but limited in its capabilities. Most of the information is stored in the form of documents either offline or online. Those are real facts which are in the natural language format, which are to be interpreted to extract knowledge. For that we need to perform a lot of operations at extremely high speed, which may need approaching a lot of space like terabytes. Most of the information exists on internet which continues to increase for each day. The internet is a service provided for the exchange of services throughout the world. Internet as an infrastructure provides website hosting, file transfer, electronic mail, etc. Website hosting deals with website that belong to government, academic, industrial, social networking, blogging, shopping, etc, that provides information regarding that organization. Electronic mail is for the purpose of communication in the form of digital medium through some domain services. File transfer refers to transmitting files over a network. Fundamentally, all these kind of services work or function using some communication protocols for sending, receiving, navigating, retaining, exchanging information over a digital medium. These websites are developed using programming languages that make the plain text to appear on the user's display terminal or web browsers. There is some kind of formatted instructions or syntax that is interpreted by a computer so that any user can be able to get the required content.

From Fig.1, interpretation of data, information, knowledge by a computer and human mind is understandable. These are closely related by slightest changes in their roles or their occurrences. Data is a set of values that can be either quantitative or qualitative type of individuals that collectively claims as information. For example, "0" and "1" are used to represent data for a computer at a low level to store and manipulate. While program execution, it is interpreted as instructions. Then information, which can be derived either from data or using knowledge which depends on the circumstances. Information resolves uncertainty of an event in any moment of time, sometimes it stands as a cause of communication. In modern era information, transforming into knowledge is a critical thing, for this purpose information is captured, generated, processed, transmitted, presented, stored. For any question posed in natural language information is the solution that exists. Information resolves uncertainty, sometimes information is stored as records for evidence, as semiotics in terms of signs. Knowledge can be referred as an implicit understanding of the theory or explicit way of dealing with facts, skills, which can be more or less formal or systematic.

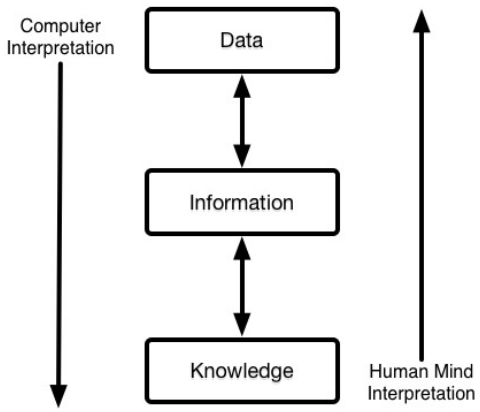


Fig.1:Interpretation of Data-Information-Knowledge

2. Related Work

Computers are considered as a convergence point of data, information and knowledge. Information system are those composed of people and the computers that process or interprets information which are helpful in almost all domains. Data acts as bridge between hardware and people. For specific reasons programs are allowed to-do machine-readable instructions on hardware of system to produce useful information from data, which is called as Software. For the development of such information system need to have a strong fundamentals to be laid to deal with all kinds of data, that is why information systems are modelled depending on kind of data that system. There are three kinds of data, they are:

- **Structured Data—**which deals with organising and relating data elements between one another. It is also called as Data Model, which explicitly determines the structure of data. These kind of data are managed through programming languages designed specifically for performing all kind of operations on it.
- **Semi-Structured Data—**a self-describing structure with tag as separator between elements and represent ordered records and within data.

- **Unstructured Data—**source that does not have pre-defined data model, typically text-heavy and contain ambiguities which cant be dealt with traditional programs.

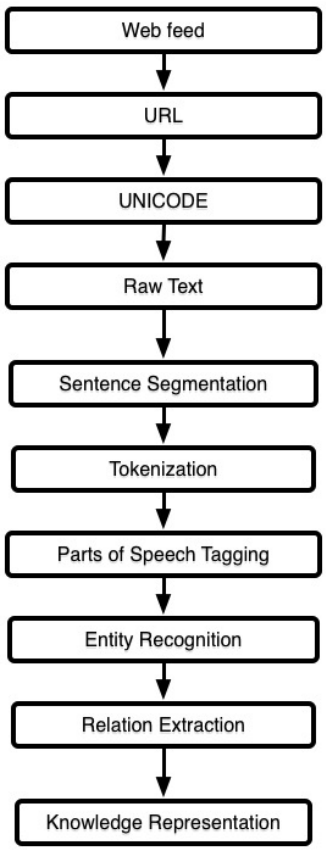
Dealing with large amounts of structured and unstructured data is an important task of Information System. The structural integration and manipulation aspects of the data stored are described by data structure. From [4], most of the existing systems are build using only structure data and others with either semi-structured data or unstructured data. Fundamentally a structured data is built by using data model, these implicitly use metadata. Metadata is defined as 'data about data', these are of two type: structural and descriptive. The data that deals with the containers of data is called Structural Metadata where the data content or individual instances of content is called Descriptive Metadata. With the help of such metadata organize electronic resources, discovery of relevant information. Metadata registry or repository is a database, where storing, manipulation of metadata is performed.

From [5], metadata is in the form of tag. A Tag is a non-hierarchical keyword assigned to a piece of information. It helps to describe an item and allows itself to find while browsing or searching. These are popular in websites and are generally chosen informally and personally by viewer depending on system. Tag play an important role in exploring records with the use of keywords in computer based search. Tagging gained popularity with the increase of social networking, bookmarking, photo sharing , etc. Using tags in system, an item can be classified in many ways with no wrong choice, instead one item belong to one category can have other tags too. Combined structural hierarchal or flat tagging can help in retrieval of information. There are some special kind of tags, they are: Triple Tags, Hashtags, knowledge tags. Triple Tags or Machine tag are special type which uses semantic information about the tag, which makes it easier or meaningful for interpretation by a computer program. Microformat is one kind of metadata that allows to add data on-page in a way users cannot see but computers can readily access. Metadata of webpages are added to search engines. Hashtags is a kind of metadata tag prefix #, sometimes known as "hash" symbol, largely used in microblogging,social networking services. Knowledge tag is of type that defines information source. These capture knowledge in the form of descriptions, categories, classification, semantics, annotations, references that are collected in tag profile.

Fig.2: Approach for Representing Knowledge

3. Proposed Work

Computing with natural language is not that easy, to do such kind, the approach has been a tough because computers can't think or act like humans. Natural languages are great source of knowledge as of humans because it contain meaning, context, background knowledge and that is why it plays a major role in communication. From [6], anticipating natural language is possible with their semantics. The semantics of natural language are its words, phrases(noun, verb) and their grammatical syntax. Most of the natural languages are based on the grammar, which organize words into sentences. The basic structure of sentence is determined by the position of the word, which is again dependent on the parts of speech. For example, in english language basic part of sentence is Subject, Object and Verb. Subject is usually noun phrase i.e. either a noun combined with a determinant or predicate. Object receives the action of verb, Verb tells what subject does. The primitives of english language are words,



collectively makes a sentence and set of sentences make a paragraph, information is to be extracted from those paragraphs using Natural Language Processing. Prior to language semantics, dealing with wide range of documents or accessing text is important.

From [7-9], knowledge representation in natural language processing can be defined as extracting useful information from natural language by using mathematical techniques and represent the knowledge. As illustrated in Fig.2 the process of representing knowledge from web feed to knowledge representation is followed by a pipeline and is explained below in detail:

- **Web feed:** The option for which kind of knowledge to be acquired by selected feed among the RSS feed of real-time. The feed of choice contains URL, title, timestamp or published time of article which are to be extracted and saved for further use.
- **URL:** It is an acronym for Uniform Resource Locater which contain protocol type and resource name. Most of time those protocol is HTTP and resource name is a web page i.e. HTML page. Through the type of protocol the web page is acquired and further processed.
- **UNICODE:** From [12], it is clear about how to manipulate strings that are acquired from web. With the help of UNICODE each and every character sequences are processed in which web page containing markup tags, empty space are removed.
- **Raw Text:** To derive only content that is important for natural language processing, removing of unnecessary code by slicing the above and below of content and cleaning that content from unwanted tags if in case leads to raw text. In this raw text, word level operations of spelling corrections, space handling between symbols are performed with respect to Word-Net [10].
- **Sentence Segmentation:** In this phase huge amount of string type is segmented by using “.”(full stop) as delimiter. Before that strings are normalized by lowering and splitting the content by delimiter.
- **Tokenization:** Words are considered as tokens and from each segmented sentence, words are segmented by using space as delimiter.
- **Parts of Speech Tagging:** The process of tagging tokens depending on their parts of speech or their position of occurrence in a sentence. Most of the words can be observed in [10,11], with that parts of speech tagging is performed without errors.
- **Entity Recognition:** This is a sub-process of information extraction where identification, classification and exemption of elements i.e. based on parts of speech. Chunking and Chinking are techniques in which each selected or defined grammar of segmented sentences with multi token sequences are parsed is a chunk and those which are exempted are chink. This is done depending on specific type of requirement or application
- **Relation Extraction:** With the identification of entities in place extracting relations that exist between them is easy. For english language basic structure of sentence is subject, verb and object are dominant sequences that relation is to be extracted for each sentence or extracting the relative noun phrase with its verb phrase is possible.
- **Knowledge Representation:** The relations that are extracted are represented with its individual parts of relationships either in the form of noun-verb phrase or in subject-verb-object form. By seeing in such an abstract manner it can be easy to understand depending on relationships for computers and can be utilized further.

With the extraction of semantic relationship of natural language and representing it with their respective relations for any given document which can be either from web or any other format like electronic books, ms word, pdf. In the above pipeline it is very important to handle with unicode [12] in digital medium. With [13,14] care taken on unicode it is easy to deal with any kind of character occurrences while processing raw text from internet to information extraction.

4. Conclusion

Finally the extraction of knowledge from unstructured data is one good way to deliver the semantic relationships exist in any given input. This method can be integrated to information extraction systems like search engines to search over the internet. Most search engines, searches data based on metadata and indexing, which always show the matched text results from ranking based algorithm. These search engines does not work on information exploring, which should be like domain independent and they don't even understand the contextual meaning that lie in the query. Exploratory search is a method followed in problem solving techniques of artificial intelligence, in which the resources are

consequently explored on a wide range (but not domain independent) of specific tasks. By integrating this type of search, there will be increase in accuracy of user search based on the contextual understanding terms that closely leads to exploratory search. In this we perform knowledge extraction from information but unable to deal with word sense disambiguation and building a graphical representation for entities based upon their relationships can be supplied as knowledge base containing facts of entities for expert systems can be the future work.

5. Reference

[1] D. A. Ferrucci, "Introduction to 'This is Watson,'" IBM Journal of Research and Development, vol. 56, no. 3.4. pp. 1:1–1:15, 2012.
[2] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. a. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building Watson: An Overview of the DeepQA Project," AI Magazine, vol. 31, no. 3. pp. 59–79, 2010.
[3] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," Natural Language Engineering, vol. 10, no. 3–4. pp. 327–348, 2004.
[4] B. Katz and B. Katz, "Annotating the World Wide Web Using Natural Language," in Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97), 1997.
[5] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," J. Web Semant., vol. 7, no. 3, pp. 154–165, 2009.
[6] A. Peñas and E. Hovy, "Semantic enrichment of text with background knowledge," in Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, 2010, pp. 15–23.
[7] L. Schubert, "Can we derive general world knowledge from texts?," in Proceedings of the second international conference on Human Language Technology Research - 2002, 2002, pp. 24–27.
[8] J. Pustejovsky and B. Boguraev, "Lexical knowledge representation and natural language processing," Artificial Intelligence, vol. 63, no. 1–2. pp. 193–223, 1993.
[9] S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text," Mach. Learn., vol. 34, no. 1–3, pp. 233–272, 1999.
[10] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," Int. J. Lexicogr., vol. 3, pp. 235–244, 1990.
[11] K. K. Schuler, "VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon," 2005.
[12] N. a. John, "The construction of the multilingual internet: Unicode, Hebrew, and globalization," J. Comput. Commun., vol. 18, no. 3, pp. 321–338, 2013.
[13] J. Stewart and J. Uckelman, "Unicode search of dirty data, or: How i learned to stop worrying and love unicode technical standard #18," in Digital Investigation, 2013, vol. 10, no. SUPPL.
[14] Joel Spolsky, The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!), http://www.joelonsoftware.com/articles/Unicode.html

Gollapalli Nithin Kumar received B.Tech Degree in Computer Science Engineering from KSRM College of Engineering, Kadapa, affiliated to Sri Venkateshwara University, Tirupati, A.P, India during 2008 to 2012. Currently pursuing M.Tech in Artificial Intelligence from JNTUA College of Engineering, Anantapuramu, A.P, India, during 2013 to 2015 batch. His areas of interest are Artificial Intelligence, Machine Learning, Natural Language Processing and Computer Networks.