INNOVATIVE FEATURE SELECTION FOR EFFECTIVE CONTEXT RESOLUTION USING NATURAL LANGUAGE QUERY INTERFACE

Dr. Amisha Shingala, Dr. Priti Sajja

Abstract: Any system that supports human interaction through natural language has high utility and ease of use. The challenge in natural language arises due to difficulty in correct interpretation, disambiguation and context resolution. Use of natural language for information retrieval and other related activities enhances effectiveness of the process and provides greater flexibility to the users in terms of document access. To do so, use of a feature vector with respect to different perspectives in addition to metadata is proposed. The work presented here encompasses a generic architecture of context resolution and categorization of document through use of natural language to achieve the intended goal. The architecture encompasses various document indices along with methodology for lexicon analysis. It also uses metadata. The proposed document features (indices) along with lexical analysis will help in correctly determining the context through the limited query keywords. The architecture is domain independent and can be used for various applications in vernacular languages. To demonstrate the application of the architecture and its methodology, necessary discussion is also included in this paper with required technical details.

Keywords

Context Resolution, Lexical Analysis, Natural language Interface, Document Features, Text Categorization

I. INTRODUCTION

Large number of techniques have been developed for retrieval of information from unstructured documents / text such as rule based and statistical but when it comes to human interaction then it becomes a challenging task. Any system that supports interaction with human being through natural language will have high utility and ease of use. Natural language query needs to use some form of query understanding in order to create easily machine-readable document / text to extract the relevant sentences. For example, when person types “Mrs. Monalisa” as input text, machine cannot attempt to translate Mrs to appropriate personal title in target language. Similarly, information retrieval should not attempt to expand “Monalisa” to all morphological variants or to suggest synonyms [12]. This needs to format some named entity recogniser rules for categorization of text / documents before processing to syntactic analysis. Moreover, for syntactic analysis, many users use a Stanford parser, which provides grammatical relationships between words in a sentence to extract textual relationships which also gives ambiguity in the results.

An ambiguity can be raised in order to identify proper noun using any available parser. For example, if we use Stanford parser, and input the sentence such as – 'Mrs. Pooja studies in M.C.A. department' which does not gives proper parsing as demonstrated in section 3. Thus structural ambiguity exists with respect to possessive pronoun which indicates relationship between two names or may constitute a component of single name [10]. All the ambiguities can be taken care of if proper names, organisation names and date index are identified correctly. In order to overcome the above mentioned limitations, we propose a generic architecture of context resolution and lexical analysis.

In rest of the paper, besides the above mentioned architecture, suggested methodology for document indices and grammar for lexical analysis will help in determining correct context through query language or query keywords for document / text classification. Working of the architecture within a selected domain is also presented in the paper to show application of the proposed research work.

II. ARCHITECTURE FOR CONTEXT RESOLUTION AND CATEGORIZATION OF DOCUMENTS

The architecture of enhancing tokenisation and categorisation of document using context resolution is shown in Fig. 1.

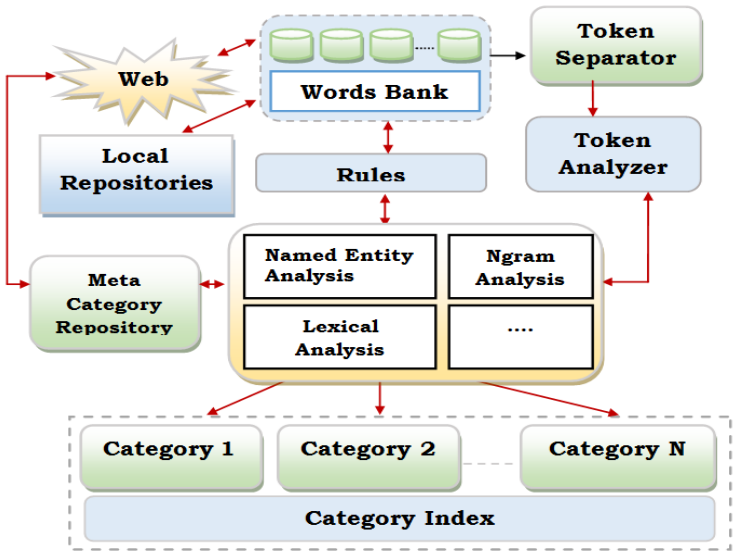

Fig. 1. Architecture of Context Resolution and Lexical Analysis

The figure represents the architecture of context resolution and lexical analysis, which is divided into two parts / sections. Part I represents how user can input the document from web or local repository, then context from document is to be extracted which uses bag of words, name index, organisation index, date index and the metadata with proper rule sets, which is shown as words bank in the architecture and the methodology is discussed in section 3. After successful completion of Part I, the pre-processed document is further processed for Part – II. It includes token analyser and token separator methodology and grammar of lexical analysis of the document is presented which is described in section 4.

III. METHODOLOGY FOR FEATURE SELECTION FROM DOCUMENT INDICES

We here discuss the three document indices based on rule set method.

Rule 1: Person Name: Document Index: Let D be the document; a Person Name Document (PN^D) is: PN^D belongs to (D, Pn1, Pn2 … Pnr), Pnr belongs to {1..r} being Person-Name and D the related document. For Example: For Proper Noun called PN-Do-Io we have eight Documents' indices:

Pn1= Metadata of contextual words includes books, author of, co-author, worked, state, city, country, university, college, school, island-of, hero, hospital, born, establish, started, saints, founded, chairman-of , director etc.

Pn2= Set of capitalized word include a set of letters followed by (.), followed by mostly one (rarely two) capitalized words.

Pn3= The words which can be immediately preceding to potential name such as Mr, Shri, Prof, Dr, Mrs, Army, Air force, Navy rank, Justice, H.H, Master, St, Ratna, Padmashri , Sir, His Excellence, Rev, Lord, Swami, Brother, Sister, etc.

Pn4= One of the capitalized words appears subsequently [3].

Pn5= Set of words or one of capitalized words appear at the beginning of a sentence.

Pn6= The Preposition such as by, of, friend, colleagues, to, co-author, with, men, persons, emperor, men like, sage, as, etc.

Pn7= Word immediately after the capitalized word(s) (ie. The post-position) is belongs to set {said, told} [3]

Pn8= An apostrophe's ('s) to a capitalized word.

Rule 2. Place/Institute/Organization Name: Document Index: Let D be the document; a Place/Institute/Organization Name-Document (PO^D) is: PO^D belongs to (D, Po1, Po2…Por), Por belongs to {1..r} being Place/Institute/Organization Name and D the related document.For Example: for Place/Institute/Organization called PO-Do-Io we have two Document indices:

Po1= A Preposition such as 'of, from, it, to' comes immediately after a Name.

Po2= Possible set of preposition for potential Place or Organisation such as 'Corp, inc, co.' etc.

Rule 3. Date: Document Index: Let D be the document; a Date - Document (DT^D) is: DT^D belongs to (D, dt1, dt2…dtr), dtr belongs to {1.r} being Date and D the related document.For Example: for Date called DT-Do-Io we have one Document index:

DT1= Words like {Century, decade, AD., BC. during, before, after, until} then probably it's a year.

The methodology, which was discussed above, will helpful for the context resolution when Stanford parser fails to resolve as discussed in section 1. The experimental result is shown in following figures. Fig. 2 represent Ambiguity in abbreviation resolution where rule based method will work.

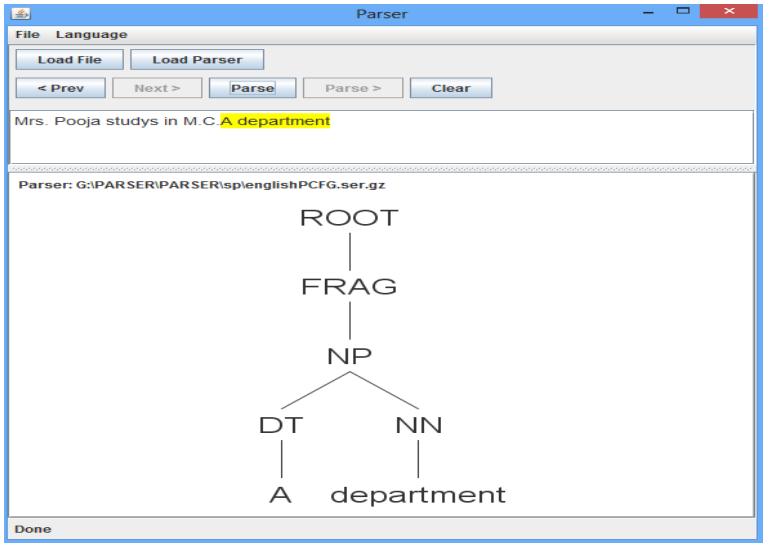

Fig. 2. Ambiguity in Abbreviation Resolution

Fig. 3 shows preposition rule for parser where rules for place/organisation can be identified.

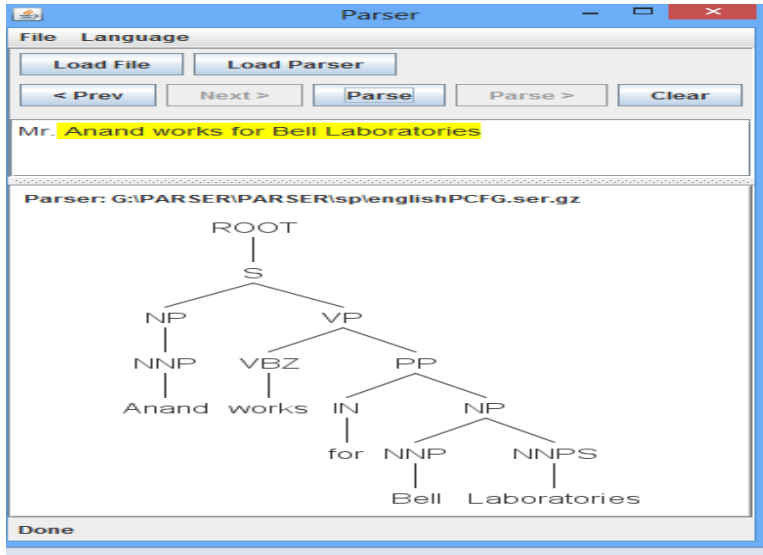

Fig. 3. Preposition Rule for Parser

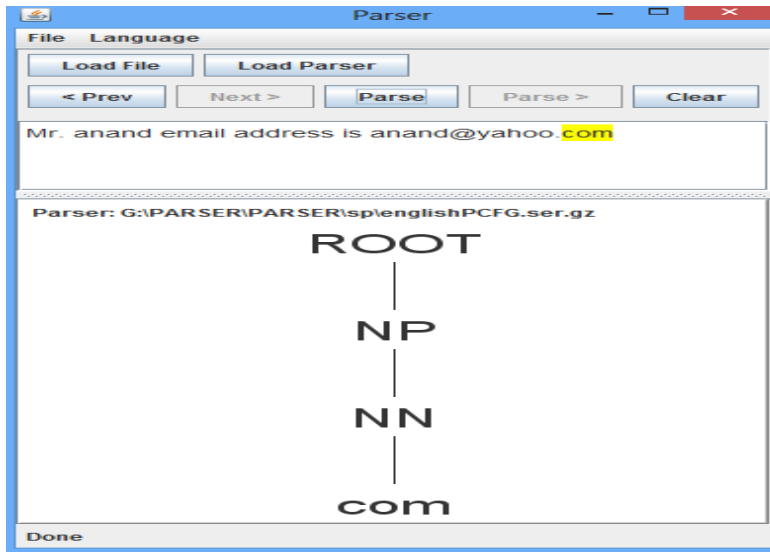

Fig. 4. Ambiguity in Email resolution

Fig. 4 shows Ambiguity in email resolution that indicates that context resolution rules are analysed before proceeding to next step of natural language query processing.

IV. GRAMMAR FOR LEXICON ANALYSIS

The token representation grammar helps in identification of lexical analysis before proceeding to syntactical analysis. The document/ text is divided into smaller fragments called tokens. Tokenization algorithm starts with basic sentence segmentation separating strings into basic tokens. Through grammar rules each identified token is categorized into different types of token representation.

<select token> := <tokenise> <symbol><select category>
<tokenise> := token (any TOKENISE-ALGO)

<select category> := <alphabets>| <Numerical>|<Special>

<alphabets> := A..Z / a..z

<Numerical> := 0..9

<Special> := <capital alphabet><fullstop> | <Numerical><fullstop> | <alphabet><quotes> | <alphabet><any | symbol> <alphabet>

<fullstop> :=.

<quotes> :=””

<symbol> := [?] [.][!][@][$][-][_]

The token representation grammar given above helps in determining the most sequence of word likely to come next, given the prior context is required. Such word prediction is useful in communication where people can't speak due to some form of disability. Sometimes, word prediction is also helpful to user when choosing menu by writing initial letter.

V. CONCLUSION

The paper contributes a generic domain independent architecture and designs various document features indices along with methods of lexical analysis. The proposed research work is tested for ambiguity. It has been observed that, around most of cases, the system could correctly interpret and process the text. With little modification, in future, the architecture may be used for spoken discourse. Further, the proposed work provides rich way to handle text processing to domain specific applications and generation of various indexed such as preparation of index at end of books and other forms of documents.

REFERENCES

[1] Alessandro Moschitti Natural language processing and automated text categorization, Ph.D.thesis, (2003).

[2] Andrei Mikheev, Periods, Capitalised Words, Article published at University of Edinburgh, association for computer linguistics, (2002).

[3] Amisha Shingala, Dr. Paresh Virparia, Enriching Document Features for Effective Information Retrieval using Natural Language Query Interface, International Journal of IT, Engineering and Applied Sciences Research, (2012).

[4] Barzilay and Elhadad, R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid,(1997)

[5] Claude de Loupy, Eric Crestan, Elise Lamaire, Proper Noun Thesaurus for Document Retrieval and Question Answering, project by French government, (2001).

[6] David Nadeau & Santoshi Sekine, A survey of named entity recognisation and classification, National Research council Canada/New York Univeristy, (2007)

[7] Joanna Rabiega, Syntactic Structure of Polish Proper Names of Places, ICS PAS Warsaw, Porland, (2006)

[8] Marcus Hassler & G.unther Flied, Text Preparation through extended tokenization, Article published at University Klangenfurt, (2006).

[9] Marie-Catherine de Marneffe, Christopher D. Manning, "The Stanford typed dependencies manual" in Revised for Stanford Parser v1.6.2, ,(2010).

[10] Nina Wacholder et.al, Disambiguation of proper names in text in the proceedings of Natural Language processing conference, Washinton D.C, (1997)

[11] Ratinov, L. and Roth, D, Design challenges and misconceptions in named entity recognition (2009).

[12] Wacholder N, Y.Ravin, Retrieving information from full text using linguistic knowledge, in proceedings of fifteen national online meeting, New York, (1994).

AUTHORS' PROFILE

Dr. Amisha H. Shingala is working as an Assistant Professor in Dept. of MCA at SVIT VASAD having more than 20 years of experience. She is an MCA, M.Phil and completed her Ph.D in 2014 from S.P.U. Her area of interest: includes NLP, Machine Learning and Data Mining, Data Analytics.

Dr. Priti S. Sajja been working as a Professor at the Post Graduate Department of Computer Science, Sardar Patel University, India. Her research interests include knowledge-based systems, soft computing, multi-agent systems, and software engineering. She has produced more than 180 publications in books, book chapters, journals, and in the proceedings of national and international conferences out of which five publications have won best research paper awards.