Information Retrieval in Malayalam Using Natural Language Processing

Merlin Rajan, Rinku T.S, Varunakshi Bhojane

Abstract- This paper explains the information retrieval using natural language processing for Malayalam language in these basic things such as types of information retrieval, the relation of natural language processing with information retrieval.

Index Terms- Clustering, Information Retrieval, Malayalam, Parsing, Tagged Text Parser, Word Suffix Trimmer, Text Skimming.

1 INTRODUCTION

Malayalam is a language spoken in India, predominantly in the state of Kerala. It is one of the 22 scheduled languages of India and was designated a Classical Language in India in 2013. The earliest script used to write Malayalam was the Vatteluttu script, and later the Kolezhuttu, which derived from it. The oldest literary works in Malayalam, distinct from the Tamil tradition, are the *Paattus*, folk songs, dated from between the 9th and 11th centuries. Grantha script letters were adopted to write Sanskrit loanwords, which resulted in the modern Malayalam script.

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Information retrieval used to be an activity that only a few people engaged in: reference librarians, para legals, and similar professional searchers.

Merlin Rajan
Information Technology Departmaent, Mumbai University
Email-Id: merlintharakan@gmail.com

Rinku T.S
Information Technology Department, Mumbai University
Email-Id: rinkusree@gamil.com

Varunakshi Bhojane
Computer Department, Mumbai University
Email-Id: varunakshi_k@yahoo.com

Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching. Information retrieval task is to select documents from a database in response to a user's query, and rank these documents according to relevance. Widespread belief that automated NLP may not be suitable in IR.

These difficulties included inefficiency, limited coverage, and prohibitive cost of manual effort required to build lexicons and knowledge bases for each new text domain. On the other hand, while numerous experiments did not establish the usefulness of NLP, they cannot be considered conclusive because of their very limited scale .Ways to overcome the poor statistical behaviour of syntactic phrases has led to various clustering techniques that grouped synonymous or near synonymous phrases into "clusters" and replaced these by single "metaterms".

Clustering techniques were however somewhat successful in upgrading overall system performance, but their effectiveness was slowed down by frequently poor quality of syntactic analysis. Information retrieval systems can also be distinguished by the scale at which they operate:1) Web search, 2)Personal information retrieval ,3)Enterprise, institutional, and domain-specific search.

A typical information retrieval (IR) task is to select documents from a database in response to a user query, and rank these documents according to relevance. This has been usually accomplished using statistical methods (often coupled with manual encoding) that (a) select terms (words, phrases, and other units) from documents that are deemed to best represent their contents, and (b) create an inverted index file (or files) that provide and easy access to documents containing these terms. A subsequent search process will attempt to match a pre-processed user query (or queries) against term-based representations of documents in each case determining a degree of relevance between the two which depends upon the number and types of matching terms.

2 TYPES OF IR

Information retrieval (IR) system aims to retrieve relevant documents to a user query where the query is a set of keywords. CLIR involves the retrieval of documents in a language other than the query language. Since the language of query and the documents to be translated in CLIR. But this translation causes a reduction in the retrieval performance of CLIR as compared to monolingual IR system. The main reason for this reduced performance is missing specified vocabulary missing general terms and wrong translation due to ambiguity. With the start of the Internet, information retrieval became increasingly relevant and researched. Now, most people use some type of modern information retrieval system on a daily basis, whether it is Google or some specially created system for libraries. This deals with asking question in one language and retrieving documents in one or more different languages. The variants of the IR are BLIR, CLIR and MLIR [2]. CLIR deals with asking questions in one language and retrieving documents in different language. MLIR deals with asking questions in one or more languages and retrieving documents in one or more different languages.

3NATURAL LANGUAGE PROCESSING

NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language. Also called Computational Linguistics. Also concerns how computational methods can aid the understanding of human language NLP research pursues the elusive question of how we understand the meaning of a sentence or a document. Natural Language Processing (NLP) is the engineering of systems that process or analyze written or spoken natural language.

3.1 Approaches in NLP

3.1.1) A statistical approach

Statistical processing of natural language represents the classical model of information retrieval systems, and is characterised from each document's set of key words

3.1.2) A linguistic focus

This approach is based on the application of different techniques and rules that explicitly encode linguistic knowledge. The documents are analysed through different linguistic levels.

4 PARSING

Parsing or syntactic analysis is the process of analysing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar.

4.1 Fast parsing with TTP parser

TTP (Tagged Text Parser) [1], [3] is a top down English parser specifically designed for fast, reliable processing of large amounts of text. The parser operates on a tagged input, where each word has been marked with a tag indicating a syntactic category: a part of speech. Top-down parsing is a parsing strategy where one first looks at the highest level of the parse tree and works down the parse tree by using the rewriting rules of a formal grammar.

In computer science, parsing reveals the grammatical structure of linear input text, as a first step in working out its meaning. Bottom-up parsing identifies and processes the text's lowest-level small details first, before its mid-level structures, and leaving the highest-level overall structure to last. TTP is a full grammar parser, and initially, it attempts to generate a complete analysis for each sentence. However, unlike an ordinary parser, it has a built-in timer which regulates the amount of time allowed for parsing any one sentence. If a parse is not returned before the allotted time elapses, the parser enters the skip-and-fit mode in which it will try to "fit" the parse. While in the skip-and-fit mode, the parser will attempt to forcibly reduce incomplete constituents, possibly skipping portions of input in order to restart processing at a next unattempted constituent.

5. WORD SUFFIX TRIMMER

The suffix trimmer [1] performs essentially two tasks:
(1) It reduces inflected word forms to their root forms as specified in the dictionary.
(2) It converts nominalized verb forms (eg. "implementation", "storage") to the root forms of corresponding verbs (i.e., "implement", "store"). This is accomplished by removing a standard suffix, eg. "stor+age", replacing it with a standard root ending ("+e"), and checking the newly created word against the dictionary. Frequently, the performance of an information retrieval system will be improved if term groups, such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single term. In addition, the suffix stripping process will reduce the total number of terms in the system, and hence reduce the size and complexity of the data in the system, which is always advantageous. Many strategies for suffix stripping have been reported in literature. The nature of the task will vary considerably depending on whether a stem dictionary is being used, whether a suffix list is being used, and of course, on the purpose for which the suffix stripping is being done. The

rules for removing a suffix will be given in the form (condition) S1 S2. This means that if a word ends with the suffix S1 and the stem before S1 satisfies the given condition, S1 is replaced by S2.

6 TEXT SKIMMING

Test Skimming:[5] Partial parsing can be used to generate as complete as possible a. representation of the input text, while in text skimming the goal is to extract a particular piece of information or look for only new information. This strategy is practical when reading a recount of a news story that has already been read, or in skimming a text to find the answer to a particular question.

Full Parsing.:[5] If the application domain is highly constrained, and the system knowledge base is extensive, it may be possible to process every word in the text and find its place in the meaning interpretation of the input. This type of approach may be practical in reading weather reports, technical abstracts, and certain other confined applications.

Partial Parsing:[5] In most applications, it is not possible, using current natural language technology, to recognize each word in a text and account for its role in the input. Most text processing, therefore, relies on the "fail soft" approach of partially processing the input, tolerating unknown elements as much as possible but relying heavily on domain knowledge to make up for gaps in linguistic knowledge. This approach has been used, for example, in reading news stories.

7 CHALLENGES OF NLP

- Meeting the expectations of the user.
- Understanding ambiguity in natural language.
- Understanding the effect of context on meaning.
- Understanding the referents of phrases like he (avan), she (aval) , and it(athu).
- Speed and efficiency of the interface.
- Recognizing relevant data, while disregarding the irrelevant data like age, gender.

The main issues to solve are: Understanding the meaning of a single word, Understanding the meaning of that word in connection with other words in the syntax and finally understanding both those meanings in the context in which they are spoken

8 CONCLUSION

This paper thus explains the different types of information retrieval using natural language processing. Language Malayalam is a very context sensitive language and hence involves many difficulties in information retrieval. It also explains the parsing techniques such as top-bottom and bottom -up along with word suffix trimmer. Human-level natural language processing is an AI-complete problem. That is, it is equivalent to solving the central artificial intelligence problem—making computers as intelligent as people, or strong AI. NLP's future is therefore tied closely to the development of AI in general.

As natural language understanding improves, computers will be able to learn from the information online and apply what they learned in the real world. Combined with natural language generation, computers will become more and more capable of receiving and giving instructions.

In the future, humans may not need to code programs, but will dictate to a computer in a human natural language, and the computer will understand and act upon the instructions.

REFERENCES

[1]Tomek Strzalkowski and Barbara Vauthey, "INFORMATION RETRIEVAL USING ROBUST NATURAL LANGUAGE PROCESSING",Courant Institute of Mathematical Sciences New York University 715 Broadway, rm. 704 New York, NY 10003 tomek@cs.nyu.edu

[2] N. Swapna1 , N.Hareen kumar2, B. Padmaja Rani3INFORMATION RETRIEVAL IN INDIAN LANGUAGES: A CASE STUDY ON CROSS-LINGUAL AND MULTI-LINGUAL, 1.Research Scholar, Department of CSE, JNTU College of Engineering, Hyderabad , AP.
2. Student of B.Tech, SRIT Affiliated to JNTUH. 3. Department of CSE, JNTU College of Engineering, Hyderabad, AP.

[3] Tomek Strzalkowski and Barbara Vauthey, FAST TEXT PROCESSING FOR INFORMATION RETRIEVAL, Courant Institute of Mathematical Sciences New York University 251 Mercer Street New York, NY 10012
{tomek, vauthey }@cs.nyu.edu.

[4] *Jagadish S. Kallimani*, K. G. Srinivasa**, Eswara Reddy B.****, SUMMARIZING NEWS PAPER ARTICLES: EXPERIMENTS WITH ONTOLOGY- BASED, CUSTOMIZED, EXTRACTIVE TEXT SUMMARY AND WORD SCORING, *Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru
Technological University, Kakinada, Andra Pradesh, India
**Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India
***Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur, Andra Pradesh, India
Emails: jsk_msrit@rediffmail.com srinivasa.kg@gmail.com eswarcsejntu@gmail.com

[5] Paul S. Jacobs and Lisa F. Rau, NATURAL LANGUAGE
TECHNIQUES FOR INTELLIGENT INFORMATION RETRIEVAL,
Paul S. Jacobs and Lisa F. Rau
Artificial Intclligcncc Program GE Researcll and Development
Center Schenectady, NY 12301 USA