Real Time Linguistic Analysis using Natural Language Processing

Jagdish Chandra Patni, Anupam Singh, Hitesh Kumar Sharma

Abstract: ***There are plenty of analysis done on linguistic communication process to develop and improve systems that square measure a lot of human, which will perceive easy directions. There are many research organizations working on different projects of NLP. This paper can function as the premise of connecting likeminded folks primarily based upon their thoughts, their interests the ideas they share. In an exceedingly layman's language this is able to work on a true time basis, currently primarily based upon the user response linguistic communication process Techniques are enforced and in real time thoughts are matched with someone of vertical having interests and thoughts just like ours so, it might operate like connecting us with a world of our own.***
This paper aims at providing associate surroundings wherever like individuals move with individuals of their kind in order that they speak solely purposeful and zilch out of context by developing a Natural Language Processing based Linguistic Analysis system.

Keywords:-Natural Language Processing, Linguistic Analysis, Human Language Technology, Data Mining, Speech Recognition, Text Generation

I. INTRODUCTION

A technique where machine can become more human and thereby reducing the distance between human being and the machine can be reduced. So in easy sense informatics makes human to speak with the machine simply. There square measure several applications developed in past few decades in informatics. Linguistic communication process (NLP) is that the application of machine-driven parsing and machine learning techniques to research customary text. Applications of NLP to requirements engineering include extraction of ontologies from a requirements specification, and use of NLP to verify the consistency and/or completion of a requirements specification Most of these are very useful in everyday life for example a machine that takes instructions by voice. NLP began within the Nineteen Fifties because the intersection of AI and linguistics. NLP was originally distinct from text data retrieval (IR), that employs extremely climbable statistics-based techniques to index and search giant volumes of text efficiently: Manning provides a superb introduction to IR. With time, however, NLP and IR have converged somewhat. Currently, NLP borrows from many, terribly various fields, requiring today's NLP researchers and developers to broaden their mental knowledge-based considerably.

Revised Manuscript Received on January 15, 2020

Dr. Jagdish Chandra Patni, Assistant Professor, Department of Computer Science , University of Petroleum and Energy Studies , Dehradun, Uttarakhand, India.

Anupam Singh, is Assistant Professor in Department of Informatics, School of Computer Science , University of Petroleum , Dehradun, Uttarakhand, India.

Dr. Hitesh Kumar Sharma, working as Assistant Professor at School of Computer Science , University of Petroleum and Energy Studies , Dehradun, Uttarakhand, India.

A. Data Mining

There is a large vary of technologies and focus areas in Human Language Technology (HLT). These embody areas like language process (NLP), Speech Recognition, artificial intelligence, Text Generation and Text Mining. during this issue, we'll target 2 of those areas: NLP and Text Mining. NLP has been around for variety of decades. it's developed numerous techniques that square measure generally lingually galvanized, i.e. text is usually syntactically parsed exploitation data from a proper descriptive linguistics and a lexicon, the ensuing data is then understood semantically and accustomed extract data concerning what was aforesaid. NLP could also be deep (parsing each a part of each sentence and making an attempt to account linguistically for each part) or shallow (parsing solely sure passages or phrases at intervals sentences or manufacturing solely restricted semantic analysis), and should even use applied mathematics means that to clear up word senses or multiple parses of identical sentence. It tends to target one document or piece of text at a time and be rather computationally costly. It includes techniques
like word stemming (removing suffixes) or a connected technique, lemmatization (replacing Associate in Nursing inflected word with its base form), multiword phrase grouping, equivalent word social newer, and uses techniques primarily developed within the fields of knowledge retrieval, statistics, and machine learning. Its aim control, part-of-speech (POS) tagging (elaborations on noun, verb, preposition etc.), word-sense elucidation, anaphora resolution. Text mining is generally isn't to grasp all or perhaps an outsized a part of what a given speaker/writer has aforesaid, however rather to extract patterns across an outsized range of documents. the only sort of Text Mining might be thought of data retrieval, additionally known as text retrieval or document retrieval, what typical search engines do. However, a lot of properly Text Mining consists of areas like automatic text classification in keeping with some fastened set of classes, text agglomeration, automatic account, extraction of topics from texts or teams of text and therefore the analysis of topic trends in text streams, whereas data retrieval and alternative varieties of text mining oftentimes create use of word stemming, a lot of subtle techniques from NLP are seldom used.

B. Linguistic Analysis

The following cluster of lines explores data extraction, largely named entity extraction, for the most part exploitation machine learning techniques instead of hand-built rules, with a varied quantity of linguistic data being employed. An exquisite job summarizing numerous contributions machine learning has created in advancing the knowledge extraction task [3].

They examine variety of knowledge extraction algorithms, each as ways for extracting helpful information in and of themselves and as precursors to (or integrated components of) numerous data processing algorithms. though they appear at an excellent style of algorithms and applications, the foremost linguistic information any of them use is POS tagging. Instance Filtering to scale back the dimensions of the coaching set for supervised classification-based learning systems for entity recognition. Grouping characters into words could be a long standing drawback in Associate in Nursingy text analysis task in Chinese (as well as in Japanese and Korean) and will be thought of an NLP technique resembling grammatical construction grouping in languages like English [4]. However, their experiments show that, whereas linguistic process helps with entity extraction in Chinese, incorporating legendary word data doesn't. It's another example of a machine learning algorithmic program for locating entities in text and, like a number of those higher than, uses POS tagging and no different information processing techniques. They use associate example of narrative text of police work reports for example their new technique. With the expansion of social media, the worth of text-based data continues to extend. it's tough to research an oversized corpus of text to get the structure at intervals the info exploitation machine ways. "Computing Machinery and Intelligence" with the statement, 'Can machines think?'" . Overall, this Alan Mathison Turing check has become a basis of linguistic communication process [11]. The essence of this project is to require a corpus of text and build a prognostic model to gift a user with a prediction of following probably word supported their input. A key side of the paper is discussion of techniques leveling accuracy and quantifiability for big knowledge sets. This paper provides the analysis selections accustomed develop that prognostic text model for a corpus of over five hundred, web log articles [10]. The resultant model exists as a web-based knowledge product that permits for customizable user coverage. to boot, the work bestowed during this project follows the tenets of duplicable analysis and every one code is accessible in associate ASCII text file repository to change readers to review the approach, reproduce the results, and collaborate to boost the model.

C. Natural Language Processing

A most vital side at the commencement of any knowledge analysis project is to grasp the matter. With the appearance of social media and blogs, the worth of text-based data continues to extend. The sensible application of extract worth from text is increasing seen in areas like click-based internet promoting, client segmentation, and sentiment analysis of Twitter and Facebook comments [1,13,14]. The matter in analyzing an oversized corpus of text is to get the structure and arrangement of words at intervals the info so as to research the corpus exploitation machine ways. The essence of this project is to require a corpus (a body) of text from numerous sources, clean and analyze that text knowledge, and build a prognostic model to gift following probably word supported the previous 2 words provided by a user [8]. User input may vary from formal, skilled communication designs to informal, short messages – additional typical in social media. Therefore, data of the info characteristics within the corpus is crucial. As a concrete example, a user could sort into their mobile device - "I would really like to". A prognostic text model would gift the foremost probably choices for what following word may well be like "eat", "go", or "have" - to call a couple of. knowledge sciences area unit progressively creating use of linguistic communication process combined with applied mathematics ways to characterize and leverage the streams of knowledge that area unit text primarily based and not inherently quantitative [6]. There area unit several techniques accessible at intervals the R programing language to figure quantitatively with text. A key side of this project is to make out that techniques best promote accuracy and quantifiability for big knowledge sets. This project provides a singular contribution. different models for predicting text area unit proprietary product used on numerous mobile platforms. This project makes all the code and algorithms accessible as open, cooperative code for others to research and improve.

D. Platform for Natural language Processing [2]

OpenNLP

The Apache OpenNLP library is machine learning based mostly toolkit for the process of tongue text. It supports the foremost common information processing tasks, like tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and grammatical relation resolution. These tasks area unit sometimes needed to make additional advanced text process services [7].

NLTK

The **Natural Language Toolkit**, or usually NLTK, could be a suite of libraries and programs for symbolic and applied mathematics tongue process (NLP) for the Python programing language. NLTK includes graphical demonstrations and sample knowledge [5].

Information retrieval (IR)

This is involved with storing, looking and retrieving info. It is a separate field at intervals technology (closer to databases), however IR depends on some information processing strategies (for example, stemming). Some current analysis and applications request to bridge the gap between IR and information processing.

Information extraction (IE)

This is involved normally with the extraction of linguistics info from text. This covers tasks like named entity recognition, grammatical relation resolution, relationship extraction, etc.

II. LITERATURE SURVEY

Today internet has become the major source of communication and connecting with likeminded people. According to a recent survey conducted by dazeinfo.com in India alone more than 300 million identify internet as their main source of communication. Although a number of means and platforms are already available there is no real means of smarty connecting the user in a way that suits their needs in the best possible way.

Thus there is dire need for platforms that can predict and analyze the need of a user based on his/her requirements and then make the best possible choice by connecting the user who meets such requirements most closely and accurately. Such a platform will not only reduce the time and effort to find likeminded people but will also make the conversation much more fruitful and conclusive.

Such a platform to connect people can be constructed using NLP. By analyzing the text fragments entered by the user by concept analyzer and the comparing and matching them with those of other users we can connect a user or a group that whose concepts are closely related to each [9].

For the last decade, NLP and Linguistic Analysis have been a hot topic for a lot research and analysis. Due to large and varied applications of NLP can be in almost all major sectors of industry and is not limited to IT.

NLP is used in various automatic summarization applications. Various search engines including GOOGLE and Bing use such applications. It is also actively being used in news and broadcasting to shorten and summarize large articles using automatic summarization applications.

Other application include intelligent personal assistant and knowledge navigator for smartphones like Speech Interpretation and Recognition Interface or SIRI and Google.

An engine developed by Wolfram that answer the questions automatically; It is an online tool that answer the question by parsing the question and then transfer for computing.It also crates the knowledge base during the execution of the questions that will help for further searches.

Another popular tool “Mathematica”; that automatically convert the English sentences into source code.

III. DESIGN AND METHODOLOGY

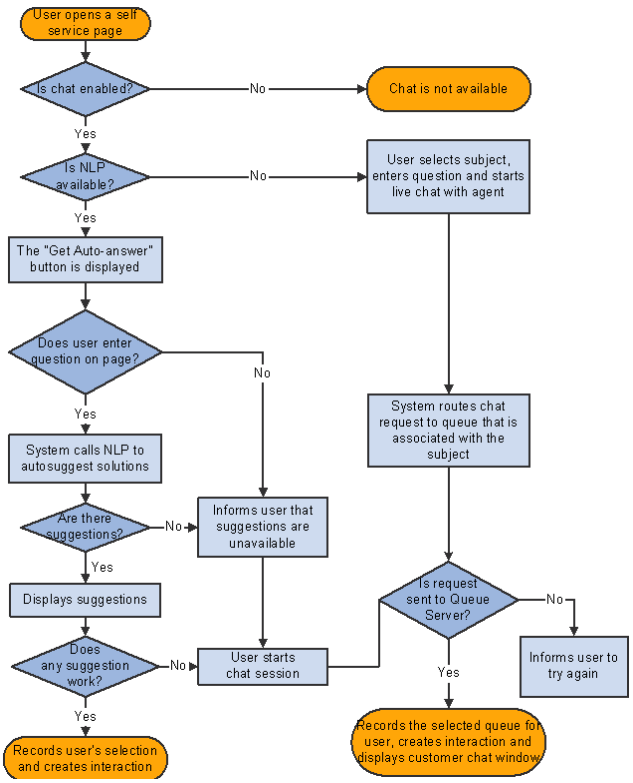

Fig 3.1 Activity Diagram using NLP

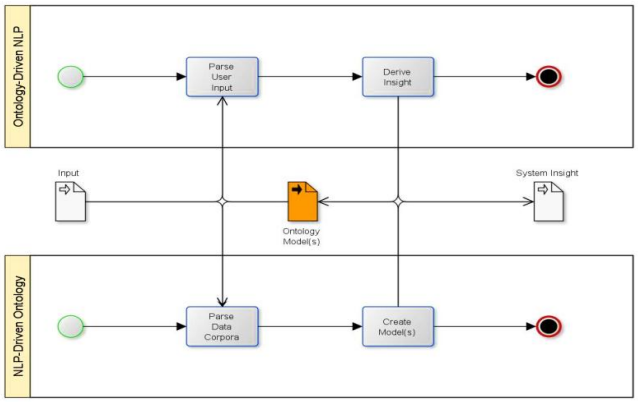

Fig 3.2 NLP Architecture

While it's true that it is possible to create very complex and accurate specifications using Spark technique, this will be an awfully time overwhelming task, moreover, it's going to need each linguistic experience, and domain experience. To facilitate this task, we will leverage the linguistic and domain experience contained within the linguistic rules and knowledge domain of a language process system to help within the creation of recent human language technology based mostly analysis system, so, we will boot-strap from AN existing system to make a replacement system that includes a richer knowledge domain victimization what we'll decision text-based idea creation, permitting a user to move with none data of human language technology

The text-based idea creation algorithmic rule consists of the subsequent seven steps.

The following is the illustration of the steps required for processing.

1. One or more test will be prompted for one or additional text fragments and this will be input for succeeding step.
2. Each fragment divided into individual words, victimization the idea analyzer.
3. The user is free to select the individual words as per their requirement from text fragment.
4. For any chosen relevant word, user will choose any synonyms, hyponyms, and hyponyms offered in Wordnet (or will mechanically embrace them).
5. A set of ideas from the user taken and it will be run on every fragment and back the results after successful search.
6. After successful match,it is important to decideeither continue with the idea or removed.
   i) Identify the relevant words.
   ii) On relevant words interpretation will be applied.
7. Build an inventory of all such relevant words that we get from previous step.

IV. IMPLEMENTATION

This paper aims at providing an environment where likeminded people interact with people of their kind so that they talk only meaningful and nothing out of context. We aim to provide an ecosystem wherein people of similar interests are set up based upon user input, be it voice input or text input which is then processed using NLP and on the other hand people having similar sort of interests are matched into the similar data sets so as to match them with a person of their origin.

Following are the steps of implementation on a web application:

1. Front-end Part: Creating a simple web-application using with user login system, registration system and area for posting the current status or interest of the person. The same is implemented using HTML, Bootstrap, JavaScript libraries, etc.
2. Back-end Part: This is the main part where our Natural Language Processing logic is implemented.
   - At the low level, various modules of NLP are used to fetch out/extract useful information from the text posted by any user. The techniques of NLP used for this purpose are:
     - ➔ Word Tokenization
     - ➔ Sentence/String Tokenization
     - ➔ Chunking
     - ➔ Chinking, etc
     - ➔ Part Of Speech Identification
   - Further a training dataset is being created which trains the system to extract wht part of the text user posts so that it can be compared with other users having similar entry in their database columns.
   - Different modules of NLP and finally that single most suitable word analyze anything, which is posted by the user, or combinations of words are extracted and saved in the MySQL database in the column containing the latest interest or status entry.
   - Finally when the user clicks on "Search people with similar interests button", if there I some user in the database with similar interest both the users are notified and the profiles of both the users are shared with each of the users. In this way the users are connected.

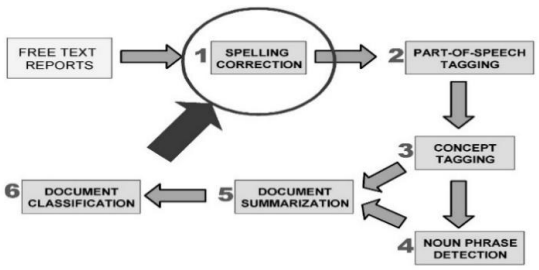

Fig 4.1 Process State Diagram

```
HTTP/1.1 200 OK
Server: nginx/1.0.4
Date: Thu, 06 Oct 2011 16:14:01 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
X-Powered-By: PHP/5.3.6

25
This is the data in the first chunk

1C
and this is the second one

3
con
8
sequence
0
```

Fig 4.2 Text Chunking

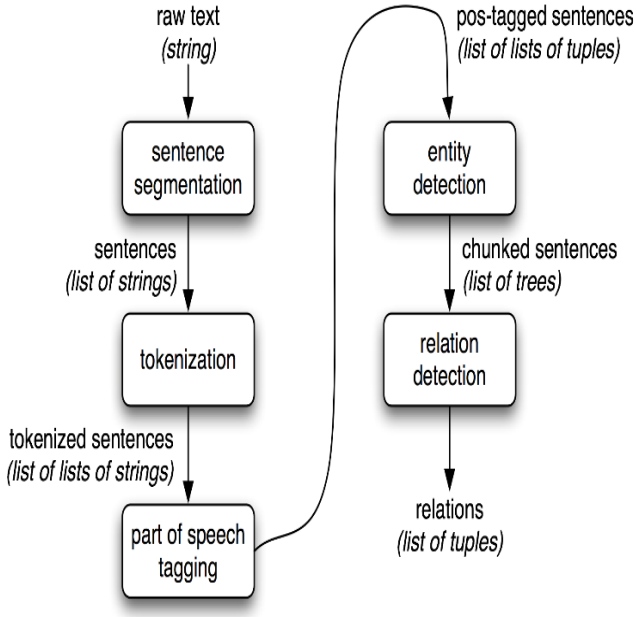

Fig 4.3Chinking Process

V. EXPERIMENT AND RESULT

It is going too trained on various parameters. The text from the paragraph cleaned executed, and replaced each digit by "JH", and others occurred less than 8 replaced by "JAH".

Table 5.1: Training data sizes

Paragraphs	T Flags	Types
345672	1345678	267890

In our work, we first train the models, then all the features incorporated in four different categories. We compare our work in the following parameters.

- We used the data 8 times by shuffling and trained the data accordingly. After shuffling the data we check the results.
- Trained the data received in the previous step by its length and occurrence.
- Based on the previous steps we analyze the word by similar meanings and models.
- In the last steps we analyze and fi the solutions for the different set of inputs.

Results shown in table 5.2. The training data and analysis shown with different factors. The given resultsalso describe the specific task support of the various activities. These results show how effectively it is working consistent in all the factors.

		Sent	Parse	POS	Data
Point1	Average	56.12	75.63	89.75	70.18
	Best	56.02	74.56	89.01	71.08
Point2	High-low	56.78	75.62	88.76	72.34
	Low-high	56.23	74.86	86.98	71.89
Point3	Performance	56.78	76.84	87.89	73.01
Point4	Nature	56.23	76.53	88.78	70.52
	Build	55.98	76.51	86.78	71.76

VI.CONCLUSION

A Natural Language Processing based mostly analysis was disbursed that will analyze every sentence entered by user into a java program such it may well be used as a file and method the desired knowledge from it therefore} reason the info and if the user says so, the method would be recurrent for every word in a very sentence.

It shows that employing a language model will improve the quantity of search and understanding;however,this is not conclusive that it's worthy to use complicated models rather than easy or simply heuristic methods.

It helps particularly once it is engineered on text have similar domain received from and in typewritten text; so it's most likely worthwhile to refer the text in the literature.If it autocompletes the works the least bit, it will once it will tends to different meaning of the word of a definite domain, instead of once it's robust generic information of language.

REFERENCES

1. Boyd-Graber, J. and P. Resnik ”Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation”. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 45–55, 2010.
2. Routledge Brady, H. E. and D. Collier, Rethinking social inquiry: Diverse tools, shared standards, 2010.
3. Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin “A statistical approach to machine translation. Computational linguistics 16 (2), 79–85, 2015.
4. Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 19 (2), 263–311,1993.
5. Witten, I.H., Cleary, J.G., and Darragh, J.J. "The reactive keyboard: a new technology for text entry" Converging Technologies: Proc Canadi an Information Processing Society Conference, 151, Ottawa, ON, May 2014.
6. Adaptive Predictive Text Generation and The Reactive Keyboard John J. Darragh and Ian H. Witten May 2012.
7. T. Honkela and A. Hyvärinen. Linguistic feature extraction using independent component analysis, pp 34-51, 2010.
8. P. Brown, P. de Souza, et al ClassBased n-gram Models of Natural Language, 1992.
9. Sami Virpioja “New methods for statistical natural language modeling”, M.Sc. thesis, 2006
10. Manning, C. and Schutze, H. Foundations of
11. Statistical Natural Language Processing,MIT Press, Cambridge, MA, 1999.
12. A. M. Turing Computing Machinery and Intelligence. Mind 49: 433-460.
13. Parul Mittal, Piyush Mishra, Vivek Patel , Jagdish Chandra Patni, “Comparison of Runtime Performance Optimization using Template-Metaprogramming” presented and published in International Conference on Next Generation Computing Technologies (NGCT), CCIS, volume 827 ,pp. 139-147, 30-31 October 2017, DOI https://doi.org/10.1007/978-981-10-8657-1_11.
14. Jagdish Chandra Patni, Hitesh Kumar Sharma, “ Air Quality Prediction using Artificial Neural Networks” presented and Published in International Conference on Automation, Computational and Technology Management (ICACTM-2019), IEEE, pp 568-572, 2019, DOI- 10.1109/ICACTM.2019.8776774.

AUTHORS PROFILE

Dr. Jagdish Chandra Patni, working as Assistant Professor at School of Computer Science , University of Petroleum and Energy Studies , Dehradun. He completed Ph.D. in the area of High Performance computing in the year 2016 M. He has completed M. Tech. and B. Tech. respectively in the year 2009 and 2004.. His area of research are Database Systems, High Performance computing, Software Engineering, Machine Learning. He has published more than 50 research papers in Journals/Conferences of Repute and published 3 International books, 2 book chapters. He is Guest Editor/Reviewer of various referred International journals. He has delivered 15 Keynote/Guest speech in India and abroad. He has organized more than 5 International Conferences as Chair/Secretary and 6 National Workshops as Chair/Co-Chair.

Anupam Singh, is Assistant Professor in Department of Informatics, School of Computer Science , University of Petroleum , Dehradun. He is pursuing Ph.D. from Dr. A P J Abdul Kalam Technical University Lucknow(Formerly UPTU Lucknow). He has done B. Tech. in 2004, M. Tech. in 2011 from Uttar Pradesh Technical University Lucknow. His area of interest are Formal Methods, Distributed System and Database System. He is Reviewer of many referred journals. He has evaluated many presentations in International Conference as Session Chair.

Dr. Hitesh Kumar Sharma, working as Assistant Professor at School of Computer Science , University of Petroleum and Energy Studies , Dehradun. He completed Ph.D. in the area of Software Engineering in the year 2016 M. He has completed M. Tech. in the year 2009 .His area of research are Database Systems, Big data, Software Engineering, Artificial Intelligence. He has published more than 30 research papers in Journals/Conferences of Repute. He has delivered 5 Keynote/Guest speech in India and abroad. He has organized more than 3 International Conferences as Chair/Secretary and 2 National Workshops as Chair/Co-Chair.