IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Study On Artificial Intelligence behind Natural Language Processing

Dr M Mary Sujatha
Asst Professor, Computer Science Dept
National Sanskrit University
Tirupati, India

Abstract: ***Natural Language Processing is a branch of artificial intelligence in which computational techniques are used to understand human languages in smart and useful way. Translation is a process through which historical scripts can be given to modern world. When compare with traditional language translation techniques, computational translation process makes it more convenient and can retrieve accurate results. Artificial Intelligence has significance in the area of linguistics. It helps to build a model for efficient and easier translation process. It makes scientific study of a language with its structure including grammar and syntax. Given study explores how Artificial Intelligence based computational techniques are helpful in natural language translation process by highlighting existing translation techniques in machine learning area and the process of developing new model. It also elaborates challenges in translation process.***

Keywords - Artificial Intelligence, Machine Translation, Natural Language Processing, Linguistics.

I. INTRODUCTION

Language is a medium of human communication. All human languages are natural languages, consists of both structured and unstructured data. Linguistics is the study of language in terms of morphology, syntax, phonetics and semantics. Linguistics use artificial intelligence to understand, analyze and derive meaning for human languages efficiently with the help of computational techniques [1]. In particular computational techniques are used to process and analyze large amount of natural language data and derived new version of linguistics called Natural Language Processing (NLP). Figure 1 demonstrates the relation between NLP, Machine Translation and Artificial Intelligence.

Natural Language Processing is an intersection of linguistics and artificial intelligence. From past two decades study on natural languages increased to make them available to layman with the help of machine translation techniques. It is mainly concerned with the interaction between machine and human languages. Machine learning models can be maintained at server, cloud side and also in smart phone as mobile applications to see, hear, sense and think [2]. It helps language translation both at script and dialog level. Machine translation helps humans to translate information such that their ideas, cultures and technological discussions can be shared with the world. In such a way it helps the world to unite socially culturally and technologically. NLP developers can organize and structure existing knowledge to perform certain tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition and topic segmentation. NLP technologies are powered by deep learning techniques due to larger amount of training data, faster machines, new models and algorithms with advanced capabilities and improved performance [3]. Unfortunately very less work has been done in the NLP for Indian languages like Hindi, Marathi, Bengali, Kannada, Telugu and Assame because of having structurally rich and idiomatic in nature. Artificial Intelligence researchers introduced semantic based approach to language analysis which requires ontological and lexical knowledge.

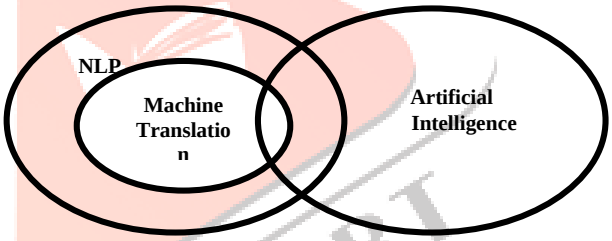


Figure 1: NLP Vs Machine Translation and Artificial Intelligence.

This paper includes Artificial Intelligence relation with linguistics. It also covers importance of natural language processing and its contributions in the study of source language and generation of target language. Furthermore, different application uses of language processing techniques explore its presence in real world. The core part of this study lists frequently used algorithms in language processing along with steps involved in translation process helps to build new translation model. The last part of this paper presents pitfalls associated with language processing.

II. EXISTING TRANSLATION TECHNIQUES

Artificial Inteligence introduced machine translation techniques, which automatically convert one natural language to another without loosing meaning of input text in translation process. There are different translation techniques based on its practice [4]. In the following grounds list of approaches were discussed, which helps in language translation process . They are broadly classified into 4 parts namely

1. Direct translation,
2. Intermediate representation,
3. Corpus based machine translation
4. Knowledge based machine translation.

1. Direct Translation- Initially word to word translation taken place with the help of dictionary look up and it failed to understand human languages in full extent. Translation is not simply a process of mapping a lexical term to other lexical term because of having underlying ambiguity in meaning of different phrases [5]. Another setback is syntax issues as it contributes to the meaning of sentence.

2. Intermediate Representation -In later period analytical description model was used in translation process. It analyzed each sentence in form of action and its auxiliaries. The action word is represented as root and auxiliary activities are represented by nominals such as naming words and adjectives [6]. The meaning of a root word includes both action and result. In other words it is linked with the phrase "to do". It indicates to find the meaning of any action it is sufficient to answer the question "What does he do" Example: "He goes" indicates "He performs an act of going", "He drinks" indicate "He performs an act of drinking", etc.

Syntactic analysis is a phase of analytical description model, where a sentence and its grammatical structure are analyzed with respect to formal grammar. Syntax gives certain rules to put words together to form components of the sentences and these components in turn made sentences. Language parsing techniques are broadly classified into three types

i. Rule based grammar driven approach
ii. Statistical based data driven approach
iii. Generalized parsers.

3. Corpus based machine translation- It requires parallel text to align against each sentence of source and target language as pair. Source sentence parsed into tokens and based on relationship between tokens semantic tree can be constructed with the help of grammar and semantic representation [7]. However it suffers from structural ambiguity, scope ambiguity and attachment ambiguity. Scope ambiguity is subject to scope redundancy on activity it got represented

Example: "Krishna is playing and watching TV" The scope of the subject "Krishna" is ambiguous to which activity it refers

Attachment ambiguity: It arises while attaching phrase or clause to a part of sentence with uncertainty.

Example: "I saw a boy with a hat", here it is not clear in association of object with subject i.e, hat existence with self or Boy is ambiguous, because most of the Indian languages have post positions instead pre positions.

4. **Knowledge based machine translation system** – It is a rule based probabilistic translation model of having semantic knowledge of source and target languages [8]. Source sentence is encoded by encoder and target sentence is predicted by a decoder. Encoder reads entire source sentence at one time and decoder uses back propagation to learn and to return translated work.

III. NATURAL LANGUAGE PROCESSING TECHNIQUES

Natural Language processing is a field of artificial intelligence. It is a two step process in which understanding of source language and generation of target human language takes place. Translation from one language to other occurs without losing original context with the help of computational techniques. NLP helps developers to organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, sentiment analysis, relationship extraction, speech recognition and topic segmentation [9].

Applications of NLP

- NLP enables the recognition and prediction of diseases based on electronic health records and patients own speech.
- Organizations can extract information from customers using sentiment analysis in sources like social media.
- Cognitive assistant works like a search engine by remembering all our personal details and help us to remind the details as personal assistant.
- NLP can classify emails and prevents spam before entering into mail box.
- It can identify fake news by detecting news sources as they are trusted or not.
- Voice driven interfaces like 'Alexa' use NLP to respond to vocal prompts, forecasts weather report, suggests best route and remind us to turn off lights in home.
- NLP is being used to track news reports, comments about possible mergers between companies. It helps financial traders.
- NLP is being used in both search and selection phases of recruitment.
- NLP helps to automate routine litigation tasks, which helps legal team to save time, work and cost.

NLP techniques are broadly classified into three types

1. Rule based systems

They highly rely on crafting domain specific rules (Example-Regular Expressions). They can be used to solve simple problems such as extracting structured data from unstructured data and fail to build models due to the complexity of human natural languages.

2. Classical machine learning techniques

Classical machine learning techniques [10] addresses problems which rule based system fail to solve. The best technique under this category is spam detection.

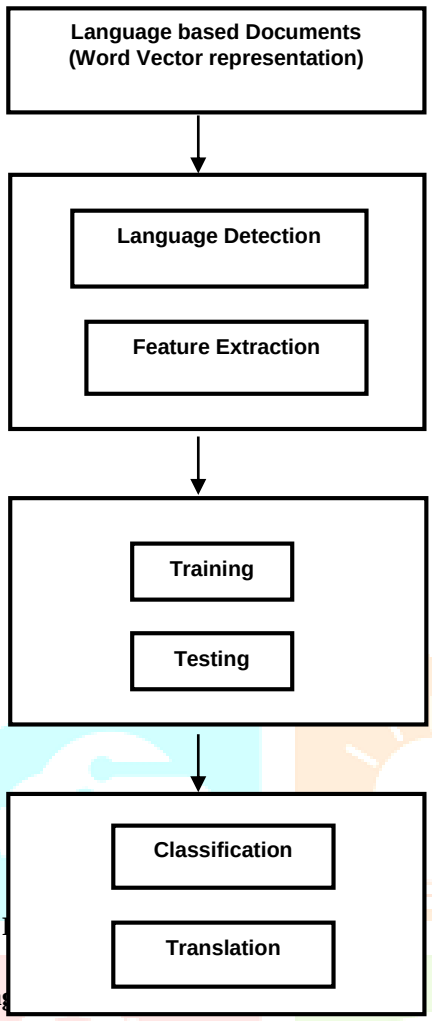
3. Deep learning techniques

Deep learning techniques are better than classical machine learning approaches as they do not have hand crafted features because they work as feature extractors in an automatic way, which help to build end to end models.

Translation Process

- Accept input data which need to be translated into vector format.
- Data Pre Processing include language detection in terms of script and feature extraction which is major task in translation process.
- Based on feature selection input data need to be classified as training data and testing data, earlier one is useful for model training and with the help of test data developed model can be tested.

- Further, data can be sub classified into different sections based on selected features and with the help of these features model can be build.



Existing

- Bag of words – It is a model that allows to count word occurrence in a piece of text. In turn a matrix can be created for all words of the document disregarding grammar and word order.
- Tokenization – In this process running text would be segmented into sentences and words. It is a task of breaking text into tokens. During segmentation process certain characters like punctuations are not considered. If the period is part of abbreviation it would be considered as a part of same token and not be removed.
- Stop words removal – In this process certain common words which give little or no value to NLP objective are filtered and excluded from the text to process. In general there is no list of universal stop words and they would be built from scratch or can consider pre-selected words. Care should be taken while dealing with certain common words like "Not", which can change relevant information and modify the context of the given sentence.

- Stemming – It is a process of slicing the end or the beginning of words with the intention of removing affixes. It reduces a word to its root level. It treats related word to its root level. It removes certain suffices like 'ing', 'ly', 'es' etc with the help of simple rule based approach. There is a chance of getting new forms for the same word. Besides its setbacks, this technique is mainly used to correct spelling errors from the tokens.

- Lemmatization – It is a process of generating target word from the source word in canonical form. The word "Lemma" means root form. It resolves words to their dictionary form. It is a process of reducing a word to its root form and groups together different forms of same word. In other words it is a process of standardizing words with similar meaning to their root form.
  Example - Go, went, gone can be considered as Go
  Though lemmatization and stemming seems to be similar, they use different approach to reach to its root form. Lemmatization is much more resource intensive task than stemming process. It requires more knowledge about the language structure than stemming approach. For instance in the process of lemmatization 'caring' turns to 'care' and stemming process might convert it as 'car'. However lemmatization is slower and takes an informed analysis.

- Topic modeling – It is a method for uncovering hidden structure in set of text or document. It discovers latent topics based on their contents by processing individual words. It assumes that each document is a combination of different topics with set of words and can spot hidden topics by unlocking the meaning of our text.
- Data vectors – It is a process of encoding text as integer i.e, numeric form to create vectors to train algorithms according to given data. It gives result of one if there is word presence in the given text and marks zero if not present.
- Features selection [12] - It uses domain knowledge to create features. This process is useful in prediction of results.
- Building model- Among existing machine learning classifiers such as SVM or Naïve Baye an ensemble method or its combination are used for results prediction. NLTK (Natural Language Tool Kit) [13] A python library that provides modules for processing text, classifying, tokenizing, stemming, tagging and parsing. iNLTK is advanced version of NLTK for Indian languages. It is an open source deep learning library built on top of pytorch platform developed in python. Currently it supports 12 Indian languages.

Conclusion

Language translation provides specific challenges for computational linguistics. First computational problem in the language translation process is in the framework of segmentation. It can be overlapped with the help of usage of parsers in translation. Human languages are culture bearings of India. Computational linguistic techniques are dominated by English and had more impact of European languages as they developed in western European countries. Most of the methods are structured to suit those European languages. Indian languages are orthographic and have inflectional complexities which are not encountered in western European languages. Most of the ancient Indian languages uses morphological analysis while doing computer translation process rather than syntactic, semantic and contextual analysis of sentences [14]. In existing machine translation systems each source language need to get translated into one or more languages before reaching to its target language which leads to a complex system.

REFERENCES

[1] Karen Sparck Jones, "Computational Linguistics: What About the Linguistics", Computer Laboratory, University of Cambridge , 2007 Association for ComputationalLinguistics Volume 33, Number 3

[2] Akshara Bharathi, Vineet Chaitanya, Rajeev Sangal "Natural Language Processing" (Chapter-5), Prentice Hall of India, https://cdn.iiit.ac.in/cdn/ltrc.iiit.ac.in/downloads/ nlpbook/nlp-panini.pdf

[3] Introduction to NLP (2016), https://algorithmia.com/blog/ introduction-natural-language-processing-nlp.

[4] Badreesh Shetty (2018) "Natural Language Processing(NLP) for Machine Learning"

[5] H. Gregory Silber and Kathleen F. McCoy "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization", Computational Linguistics 2002 28:4, 487-496

[6] Shachi Mall, Umesh Chandra Jaiswal "Survey: Machine Translation for Indian Language", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018) pp. 202-209.

[7] Tripathi, Sneha, Sarkhel, Juran Krishna "Approaches to machine translation", http://nopr.niscair.res.in/handle/ 123456789/11057, ALIS Vol.57(4) [December 2010]

[8] Nirenburg, S. Knowledge-based machine translation. *Machine Translation* **4,** 5–24 https://doi.org/10.1007/BF00367750

[9] Monika T. MakwanaȦ, Deepak C. Vegda "Survey:Natural Language Parsing For Indian Languages" https://arxiv.org/ftp/arxiv/papers/1501/1501.07005.pdf

[10] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008

[11] Diego Lopez Yse ,"Your Guide to Natural Language Processing (NLP)", https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.

[12] Cai,Jie, Luo et al, "Feature selection in machine learning: A new perspective", https://doi.org/10.1016/j.neucom.2017.11.077

[13] Alvations, stevenbird (2019) "Natural Language Toolkit"

[14] Teja et al., (2015) International Journal of Advanced Research in Computer Science and Software Engineering 5(3), pp. 596-600.