Computer Linguistics for Processing Human Language for Artificial Intelligence – Hard Applications

Mrs. Kaushika Pal, Assistant Professor, MCA Department, Sarvajanik College of Engineering and Technology, Surat, Gujarat.

Dr. Biraj V. Patel, Assistant Professor, G. H. Patel, P.G. Department of Computer Science &Technology, Sardar Patel University, V.V. Nagar, Gujarat.

Abstract:

The purpose of this research article is to explore the importance of Computer Linguistics and processing Natural Language for building Artificial Intelligence Hard-Applications, which is the need for the modern age. Human knowledge is exemplified by Language and such data on the web is increasing, generating the need of processing this huge amount of Structured and Unstructured Human Language on the World Wide Web for Information Retrieval. Artificial Intelligence applications need to extract knowledge by processing Natural Language for solving many problems to help the growth of society. The need for Critical AI-Hard Applications build by processing Human Language creates a multidisciplinary field combining literature of language and Machines to understand and process the Human Language.

Introduction

Linguistics is considered to be a scientific analysis of Human Language; Human Language is easy but at the same time complicated for various reasons. There are around 7100 different Human languages in the World; each has a set of alphabets, which are used to form a big set of words that forms vocabulary of any Human Language. The sentences, which are formed grammatically using a set of rules. The basic formation of Language is shown in figure 1. In India with 29 States and 7 Union territories, have languages belonging to 6 different categories naming as Indo-Aryan Languages, Nuristani Languages, Iranian Languages, Austro-Asiatic Languages, Dravidian Languages, and Tibeto-Burman Languages. Some Languages does not belong to any of the categories and therefore said to be Unclassified Language. Human Language can be written and spoken, and it is written or spoken by people constructing different statements to express the same meaning by different people as per their way of expressing their thoughts. Human Language forms a basis for information sharing, communication, and therefore Human beings are progressing in every field. Due to the importance of Human Language and the amount of increase in such data on the web via social networking and numerous applications necessitates the study Computer Linguistics. This field is an interdisciplinary field where literature needs to understand by machines.

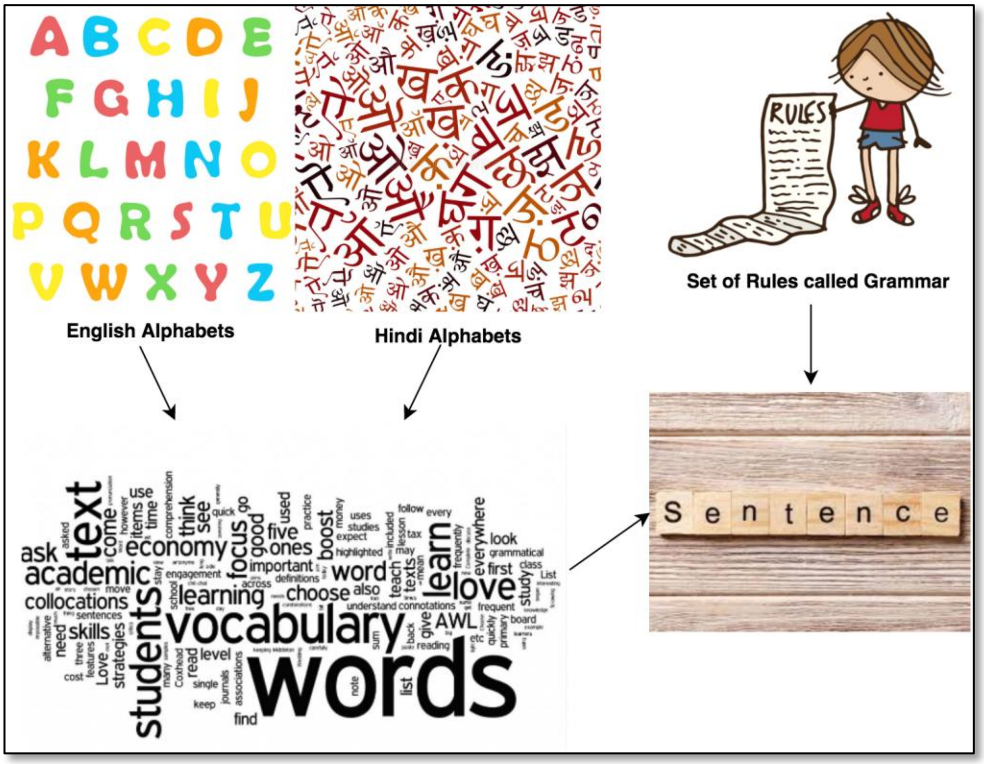

Figure 1. Formation of Any Human Language

Computer Linguistic

The goal of Computer Linguistics is to study Human Language from a scientific perspective using grammatical and semantically structured textual context to trace the elements that can be understood by machines for computational purposes. Any Language is scientifically studied using 6 different categories using different elements of the Language. Table 1 shows the details.

Table 1. Scientific Study of Elements of Language

Scientific Study	**Elements of Language**
Phonetics	Speech Sounds
Phonology	Phonemes
Morphology	Words
Syntax	Phrases and Sentences
Semantics	Literal Meaning of Phrases and Sentences
Pragmatics	Meaning in Context of discourse

Syntax and parsing, Semantic representation is done using Logicist approaches, Physiologically striving approaches, Statistical semantics approaches and all this is interpreted for making it traceable for machines.

Natural Language Processing

Natural Language Processing forms Computer Linguistics as the basis for processing Human Language to build AI-Hard applications. AI-Hard applications need to extract knowledge from the Human Language, but the entire text content represented in any language cannot be processed to decode the meaning without breaking into small pieces of information. NLP uses Computer Linguistics elements to process textual content in order to understand.

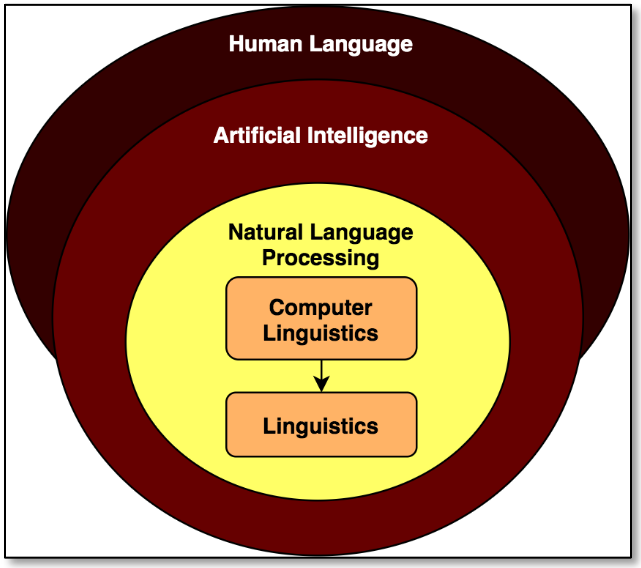

Figure 2. The Processing of Human Language for AI Applications

NLP has several constituents to do this challenging study. The constituents are Morphological Analysis, Syntactical Analysis, Semantic Analysis, and Pragmatics Analysis.

Morphological and Lexical Analysis, which breaks the statements into tokens, removes unnecessary characters, removes common words, uses lemmatization, or stemming to bring it to a level machine can extract information for AI-Hard applications.

Syntactic Analysis is done when the text is structured by using Part-of-the-Speech taggers and extract grammar-based knowledge from textual Content. Pragmatic Analysis is about extracting context in which the sentence is used for communication. It deals with reference resolution and dialog interpretation. Discourse Integration is about finding context of one statement depending on meaning of following statement.

The challenges for understanding textual content include Lexical ambiguity, Syntax ambiguity, and referential ambiguity.

Natural Language Processing deals with understanding the language and generating the language using Natural Language Understanding and Natural Language Generator components respectively.

Artificial Intelligence – Hard Applications

AI-Hard applications are build using AI components namely Natural Language Processing, Machine Learning, and Deep Learning.

NLP component of AI is responsible for techniques used for the linguistic study of textual content from a computation perspective.

Machine Learning and Deep Learning have sophisticated algorithms for Supervised Learning, Unsupervised Learning, and Semi-supervised Learning. These algorithms are used to convert various extracted knowledge from a syntactical perspective or lexical perspective into numeric form. Machine Learning techniques allow assigning weights to the knowledge using, which AI applications are builds.

The Human thought expressed on the World Wide Web from large populations may be used to serve society. These thoughts are processed to extract knowledge, which supports building AI applications based on Textual data and therefore said to be AI-Hard Applications. Figure 3 shows a general architecture followed for building Artificial Intelligence applications.

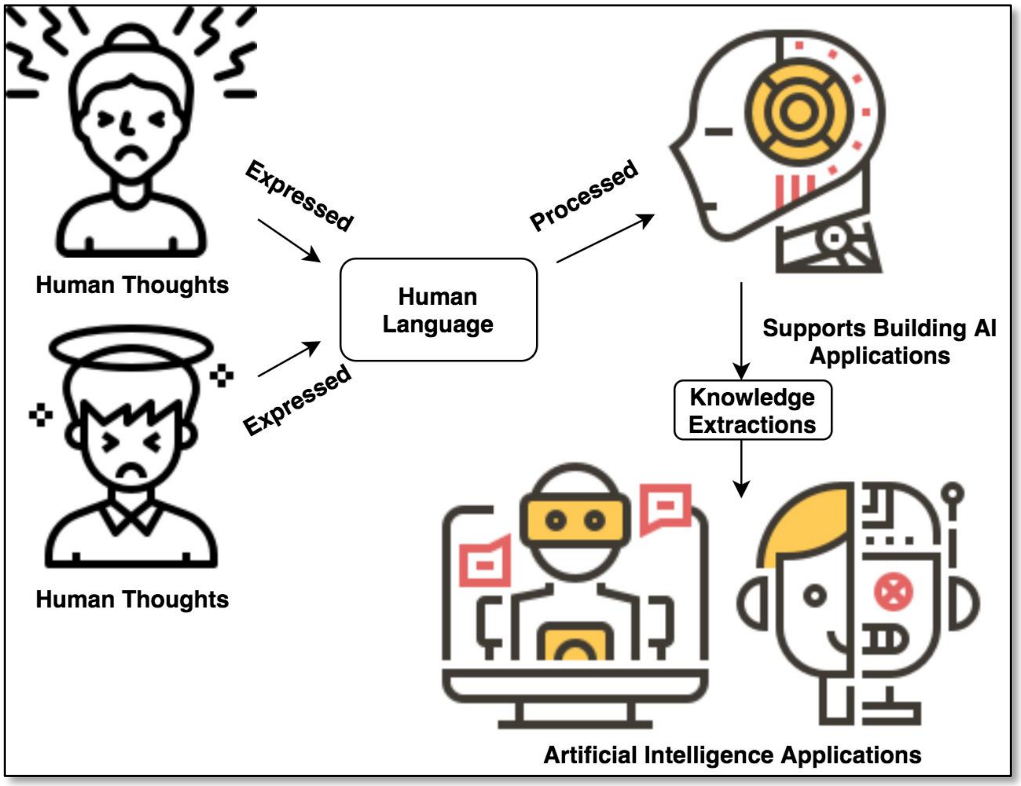

Figure 3. Human Thoughts expressed and extracted for AI Applications

There are many applications serving society, few Applications with its details are represented in table 2.

Table 2. Few examples of AI-Hard Applications

Applications	**Details**
Text Classification	This applications categories text into multiple predefined categories by understanding the textual content.
Chat bots	Chat bots are used by business to automatically answer client queries in order to increase business. This applications understand and than generates replies to clients
Text Summarization	It summarizes long textual content into small summaries. Either Extractive or Abstractive Summarization is done.
Machine Translation	Translates one language to another say Hindi to English, etc.
Emotions Identification	From written or spoken content whether a person is Sad, Happy, Excited, etc. can be predicted.
Sentiment Analysis	These applications categories review of Product or Service in Positive, Negative or Neutral for analysis to improve the product or service.
Spam Filtering	To prevent unnecessary emails in inbox this applications separates this emails in Spam category.
Topic Segmentation	The Content can divided into meaningful segments for further study.
Auto Correction	This application corrects the spellings of content. Example Word Processing

All the applications listed tries to understand the textual content and some application generates the textual content. These applications need data, which may be scattered and can be collected using web scrapping. Few applications are quite successful in English data but few applications face challenges. Researches on AI-hard applications are also building applications in various Indic Languages such as Hindi, Gujarati, Marathi, Bengali, Tamil, Punjabi, etc. each face different challenges from understanding perspective due to morphological variance.

Researchers are contributing by creating a corpus, preprocessing packages, POS tagger, Name Entity recognizer, Parser for syntactical analysis for regional and national language for all the AI-Hard applications dealing with textual data.

Table 3 shows Artificial Intelligence text classification applications on mentioned dataset of English, Arabic and Hindi Language and its performance in terms of accuracy for AI-Hard application using Natural Language Processing.

Table 3. AI-NLP, applications Results for Document Classification.

Data Set	**Algorithm Name**	**Accuracy in Percentage**
Facebook Data	Support Vector Machine	78.3%
	Naïve Bayes Classifier	77.25%
	K- Nearest Neighbor	56.42%
SMS Data	Apriori algorithm with NB	62.5%
News Group Data	Support Vector Machine	97.34%
	Naïve Bayes Classifier	95%
News Data	Neural Networks	99.28%
Unstructured Data	Support Vector Machine	97.6%
Arabic Text Documents	Naïve Bayes Classifier	68.78%
Arabic Newswire	Statistical Methods	62.7%
Arabic Literary Documents	Hybrid Approach with Tree algorithms	91%
Arabic Scientific Documents	Hybrid Approach with Tree algorithms	93%
Sentiment Analysis Movie Review Text Data	K- Nearest Neighbor	65.75%
	Naïve Bayes Classifier	74.50%
	Support Vector Machine	85%
Hindi Poetry Documents	Random Forest	56%
	K- Nearest Neighbor	52%
	Multinomial Naïve Bayes	64%
	RBF SVM	52%

The analysis of results in terms of accuracy achieved for AI – Hard applications shows that dealing with the text of any language needs improvement from Computer Linguistic perspective. Implementing solutions, which understand and reply in Human Language as Humans do, is very challenging and needs a multidisciplinary study of Literature of that Language and need usage of techniques to process that Literature with all its elements by dealing with ambiguity and removing them for robust AI-Hard applications.

Conclusion

This paper represents the importance of Human Language, which makes Computer Linguistics compulsory for AI bound applications, AI applications behave as humans think like humans and respond like humans. To achieve this complicated results machine needs to understand the human language with all its elements. The scientific study of language from the computation perspective is Computer Linguistics that forms the basis for Natural Language Processing. The NLP a branch of AI have methods Tokenize, remove special characters, numbers, common words, convert words into root form, as the machine has to deal with word-to-word basis. Machine Learning and Deep learning algorithms are used to build AI-Hard Applications. These applications can be built for any Human Language, but challenges of morphologically rich languages, ambiguity in the representation of languages make the task challenging and exciting.

References

1. Chaitanya Anne, Avdesh Mishra, Md Tamjidul Hoque, Shengru Tu, "Multiclass patent document classification", Artificial Intelligence Research, Vol. 7, No. 1, 2017, pp. 1 - 14
2. Jason D. M. Rennie, Ryan Rifkin, " Improving Multiclass Text Classification with the Support Vector Machine", Massachuseets Institute of Technology as AI Memo, 2011
3. George Yule, "The Study of Language ", Cambridge University Press, ISBN 978-1-108-44188-9, First South Asia edition 2018.
4. Tanveer Siddique & U.S. Tiwary, "Natural Language Processing and Information Retrieval", Oxford University Press, ISBN 978-0-19-569232-7, 2008.
5. Miroslav Kubat, " An Introduction to Machine Learning, Springer International Publishing", ISBN 978-3-319-63912-3, AG 2015, 2017.
6. Jalaj Thanaki, " Python Natural Language Processing – Explore NLP with Machine Learning and Deep Learning Techniques", Packt Publishing, 2017.
7. Amey K. Shet Tilve, Surabhi N. Jain (2017). "A Survey on Machine Learning Techniques for Text Classification", INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY. Vol. 6, Issue 2 (pp. 513 – 520).
8. Holzinger A., (2019), "Introduction to Machine Learning & Knowledge Extraction (MAKE)", MACHINE LEARNING & KNOWLEDGE EXTRACTION. Vol. 1 Issue. 1 (pp.1-20).
9. K Pal, B V Patel, "A Study of Current State of Work Done for Classification in Indian Languages", International Journal of Scientific Research in Science and Technology. 3(7) 403 – 407, 2017
10. K Pal, B.V.Patel, "Model for Classification of Poems in Hindi Langauge Based on Ras", Smart Systems and IoT: Innovations in Computing, Smart Innovation, Systems and Technologies 141. 2019 pp. 655-661.

11. K Pal, J Saini,“A study of current state of work and challenges in mining big data”, International Journal of Advanced Networking Applications. 73 – 76, 2014
12. Kaushika Pal, Dr. Biraj V. Patel, "Multi – Class Document Classification: Effective and Systematized Method to Categorize Documents ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Volume 7 Issue 7, pp. 118-123, January-February 2020.
13. Kaur, Jasleen and Jatinderkumar R. Saini. “Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms.” International Conference on Machine Learning and Computing (ICMLC 2017)-ACM, 2017
14. K. Pal, B. V. Patel. ““Data Classification with k – fold cross Validation and Holdout Accuracy estimation methods with 5 different Machine Learning Techniques.” 4th International Conference on Computing Methodologies and Communication. IEEE Xplore; in press Article.
15. Sbou, Ahed M. F. Al. (2019) “A survey of arabic text classification models.” International Journal of Informatics and Communication Technology Vol. 8 (pp. 25-28).
16. Raj, Jennifer S. "A Comprehensive Survey on The Comutational Intelligence Techniques and it’s Applications”. Journal of ISMAC 1, no. 03. 2019 pp. 147-159
17. Joseph, S. I. T. “Survey of Data Mining Algorithm’s for Intelligent Computing System”. Journal of trends in Computer Science and Smart technology (TCSST), 1(01), 2019 pp. 14-24.
18. Senthil Kumar B, Bhabitha Varma E. A survey on Text Categorization. International Journal of Advance Research in Computer and Communication Engineering. Vol. 5, Issue8. 2016 pp. 286-289.
19. Neha Rani, Aanchal Sharma, Sudhir Pathak (2018). “Text Classification Using Machine Learning Techniques: A Comparative Study”, International Journal on Future Revolution in Computer Science & Communication Engineering. Vol. 4, Issue. 3 (pp. 551- 555).