Current Trends and Challenges in Natural Language Processing

[1]Dr.M.Venkat Dass, [2]Dr.K.Ravi kishore
[1]Associate Professor, Dept of CSE, [2]Associate Professor, Dept of CSE,
[1]University College of Engineering (A), Osmania University, Hyderabad, India
[2]Stanley College of Engineering and Technology for Women(A),Hyderabad, India

Abstract: Current paradigm of information technology uses human-computer interaction in the form of natural language and the language we use for day-to-day communication. Natural Language Processing (NLP) has recently gained much attention for representing and analysing human language computationally. It has spread its applications in various fields such as machine translation, email spam detection, information extraction, summarization, medical, and question answering etc. The paper distinguishes various phases by discussing different levels of NLP and components of Natural Language Generation (NLG) followed by presenting the history, evolution of NLP, various applications of NLP and current trends and challenges of NLP.

Keywords: Natural Language Processing, Human Computer Interaction, Natural Language Generation

1. INTRODUCTION

Natural Language Processing (NLP) is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages. Natural language processing came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language. Since all the users may not be well-versed in machine specific language, NLP caters those users who do not have enough time to learn new languages or get perfection in it.

A language can be defined as a set of rules or a set of symbols. Symbols are combined and used for conveying information or broadcasting the information. Natural Language Processing basically can be classified into two parts i.e. Natural Language Understanding and Natural Language Generation which evolves the task to understand and generate the text.

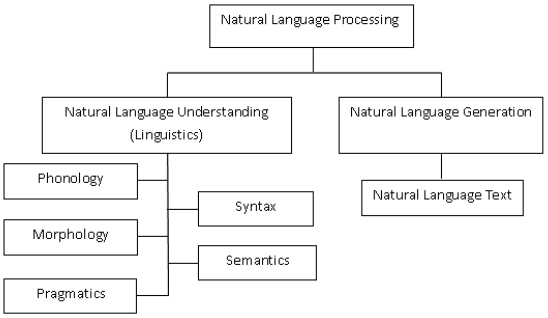

Fig 1: Classification of NLP

Linguistics is the science of language which includes Phonology that refers to sound, Morphology that refers to word formation, Syntax referring to sentence structure, Semantics referring to syntax and Pragmatics which refers to understanding.

Noah Chomsky, one of the first linguists of the twelfth century who started syntactic theories, marked a unique position in the field of theoretical linguistics because he revolutionized the area of syntax[1] which can be broadly categorized into two levels: Higher Level which includes speech recognition and Lower Level which corresponds to natural language. Few of the researched tasks of NLP are Automatic Summarization, Co-Reference Resolution, Discourse Analysis, Machine Translation, Morphological Segmentation, Named Entity Recognition, Optical Character Recognition, Part of Speech Tagging etc. Some of these tasks have direct real world applications such as Machine translation, Named entity recognition, Optical character recognition etc. Automatic summarization produces an understandable summary of a set of text and provides summaries or detailed information of text, of a known type.Co-reference resolution refers to a sentence or a large set of text that determines which words refer to the same object. Discourse analysis refers to the task of identifying the discourse structure of connected text. Machine translation refers to the automatic translation of text from one human language to another. Morphological segmentation refers to separating a

word into individual morphemes and identify the class of the morphemes. Named Entity Recognition (NER) describes a stream of text, determines which items in the text relate to proper names. Optical Character Recognition (OCR) gives an image representing printed text, which helps in determining the corresponding or related text.

The metric of NLP assess on an algorithmic system that allows for the integration of language understanding and language generation.It is even used in multilingual event detection[2].A novel modular system is proposed for cross-lingual event extraction for English, Dutch and Italian texts by using different pipelines for different languages. The system incorporates a modular set of foremost multilingual Natural Language Processing (NLP) tools.

Most of the work in Natural Language Processing is conducted by computer scientists while various other professionals like linguistics, psychologist and philosophers etc. have also shown interest. The field of Natural Language Processing is related with different theories and techniques that deal with the problem of natural language of communicating with the computers.Ambiguity is one of the major problems of natural language which is usually faced in syntactic level with further subtasks like lexical and morphological concerned with the study of words and word formation. Some of the methods proposed by researchers to remove ambiguity is by preserving ambiguity[3,4,5].

Levels of NLP

The ‘levels of language’ are one of the most explanatory methods for representing the Natural Language Processing which helps to generate the NLP text by realizing Content Planning, Sentence Planning and Surface Realization.

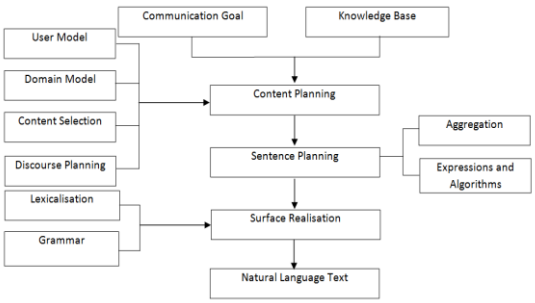

Fig 2: Phases of NLP architecture

The various important terminologies of Natural Language Processing are:-

Phonology

Phonology is the part of linguistics which refers to the systematic arrangement of sound. In 1993 Nikolai Trubetzkoy stated that Phonology that is “the study of sound pertaining to the system of language, it could be better explained as, "phonology proper is concerned with the function, behaviour and organization of sounds as linguistic items. It includes semantic use of sound to encode meaning of any human language [6].

Morphology

The different part of the word represent the smallest units of meaning known as Morphemes. Morphology which comprises of nature of words, are initiated by morphemes. The words that cannot be divided and have meaning by themselves are called Lexical morphemes (e.g.: table, chair).The words (e.g. -ed, -ing, -est, -ly, -ful) that are combined with the lexical morpheme are known as Grammatical morphemes (eg. Worked, Consulting, Smallest, Likely, Use). Those grammatical morphemes that occur in combination called bound morphemes (eg. -ed, -ing). Grammatical morphemes can be furtherdivided into bound morphemes and derivational morphemes.

Lexical

In Lexical, humans as well as NLP systems can interpret the meaning of individual words. At the lexical level, semantic representations can be replaced by the words that have one meaning. In NLP system, the nature of the representation varies according to the semantic theory deployed.

Syntactic

This level emphasis to examine the words in a sentence so as to uncover the grammatical structure of the sentence. Both grammar and parser are required in this level. The output of this level of processing is the representation of the sentence that communicate the structural dependency relationships between the words.

Semantic

Semantic processing determines the possible meanings of a sentence by pivoting on the interactions among word-level meanings in the sentence. This level of processing can incorporate the semantic disambiguation of words with multiple senses; Semantics context that most words have more than one report but that we can spot the appropriate one by looking at the rest of the sentence[8].

Pragmatic:

Pragmatic is concerned with the firm use of language in situations and utilizes nub over and above the nub of the text for understanding the goal and to explain how extra meaning is read into texts without literally being encoded in them. This requisite much world knowledge, including the understanding of intentions, plans, and goals but this aspiration requires pragmatic or world knowledge[7].

2. NATURAL LANGUAGE GENERATION

Natural Language Generation (NLG) is the process of producing phrases, sentences and paragraphs that are meaningful from an internal representation. It happens in four phases: identifying the goals, planning on how goals maybe achieved by evaluating the situation and available communicative sources and realizing the plans as a text as shown in Figure 3. It is opposite to Understanding.

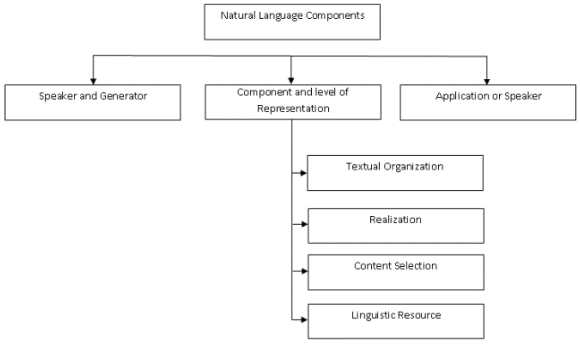

Fig 3: Components of NLG

Components of NLG are as follows:

Speaker and Generator – To generate the text we need to have a speaker or an application and a generator or a program that renders the application's intentions into fluent phrase relevant to the situation.

Components and Levels of Representation:

Content selection: Information should be selected and included in the set. Depending on how this information is parsed into representational units, parts of the units may have to be removed while some others may be added by default.

Textual Organization: The information must be textually organized according the grammar, it must be ordered both sequentially and in terms of linguistic relations like modifications. Linguistic Resources: To support the information's realization, linguistic resources must be chosen. In the end these resources will come down to choices of particular words, idioms, syntactic constructs etc.

Realization: The selected and organized resources must be realized as an actual text or voice output. The only requirement is the speaker that has to make a sense of the situation[9] .

3. HISTORY OF NLP

In late 1940s the term wasn't even in existence, but the work regarding machine translation (MT) had started. Research in this period was not completely localized. Russian and English were the dominant languages for MT, but others, like Chinese were used for MT (Booth ,1967)[10].

As early as 1960 signature work influenced by AI began, with the BASEBALL Q-A systems[12]. LUNAR[13] and Winograd SHRDLU were natural successors of these systems. The front-end projects[14] were intended to go beyond LUNAR in interfacing the large databases.

By the end of the decade the powerful general purpose sentence processors like SRI's Core Language Engine[15] and Discourse Representation Theory[16] offered a means of tackling more extended discourse within the grammatico-logical framework. Practical resources, grammars, and tools and parsers became available e.g the Alvey Natural Language Tools[17]. Some researches in NLP marked important topics for future like word sense disambiguation[23]and probabilistic networks, statistically coloured NLP, the work on the lexicon, also pointed in this direction.

Statistical language processing was a major thing in 90s[24], because this not only involves data analysts. Information extraction and automatic summarising[25] was also a point of focus.

Recent researches are mainly focused on unsupervised and semi-supervised learning algorithms.

4. RELATED WORK

Many researchers worked on NLP, building tools and systems which makes NLP what it is today. Tools like Sentiment Analyser, Parts of Speech (POS)Taggers, Chunking, Named Entity Recognitions (NER), Emotion detection, Sarcasm detectin and Semantic Role Labelling made NLP a good topic for research.

Sentiment analyser[26] works by extracting sentiments about given topic. Sentiment analysis consists of a

topic specific feature term extraction, sentiment extraction, and association by relationship analysis.

Parts of speech taggers for the languages like European languages, research is being done on making parts of speech taggers for other languages like Arabic, Hindi[28]etc. It can efficiently tag and classify words as nouns, adjectives, verbs etc. The most procedures for part of speech can work efficiently on European languages, but it won't on Asian languages or middle eastern languages. Arabic uses Support Vector Machine (SVM)[29] approach to automatically tokenize, parts of speech tag and annotate base phrases in Arabic text.

Usage of Named Entity Recognition in places such as Internet is a problem as people don't use traditional or standard English. This degrades the performance of standard natural language processing tools substantially. It improves the performance as compared to standard natural language processing tools.

Sematic Role Labelling – SRL works by giving a semantic role to a sentence. The precise arguments depend on verb frame and if there exists multiple verbs in a sentence, it might have multiple tags.

Event discovery in social media feeds , using a graphical model to analyse any social media feeds to determine whether it contains name of a person or name of a venue, place, time etc. The model operates on noisy feeds of data to extract records of events by aggregating multiple information across multiple messages, despite the noise of irrelevant noisy messages and very irregular message language, this model was able to extract records with high accuracy. However, there is some scope for improvement using broader array of features on factors.

Applications of NLP

Natural Language Processing can be applied into various areas like Machine Translation, Email Spam detection, Information Extraction, Summarization, Question Answering,Sentiment Analysis and Sarcasm Detection etc.

Machine Translation

As most of the world is online, the task of making data accessible and available to all is a great challenge. Major challenge in making data accessible is the language barrier. There are multitude of languages with different sentence structure and grammar. In this the phrases are translated from one language to another with the help of a statistical engine like Google Translate. The challenge with machine translation technologies is not directly translating words but keeping the meaning of sentences intact along with grammar and tenses.

Text Categorization

Categorization systems inputs a large flow of data like official documents, military casualty reports, market data, newswires etc. and assign them to predefined categories or indices. Another application of text categorization is email spam filters. A filtering solution that is applied to an email system uses a set of protocols to determine which of the incoming messages are spam and which are not.

Spam Filtering

It works using text categorization and in recent times, various machine learning techniques have been applied to text categorization or Anti-Spam Filtering. Using these approaches is better as classifier is learned from training data rather than making by hand.

Information Extraction

Information extraction is concerned with identifying phrases of interest of textual data. For many applications, extracting entities such as names, places, events, dates, times and prices is a powerful way of summarize the information relevant to a user's needs. There is use of hidden Markov models (HMMs) to extract the relevant fields of research papers.

Dialogue System

Perhaps the most desirable application of the future, in the systems envisioned by large providers of end user applications is Dialogue systems.

It is believed that these dialogue systems when utilizing all levels of language processing offer potential for fully automated dialog systems[7]. This could lead to produce systems that can enable robots to interact with humans in natural languages.

Medicine

NLP is applied in medicine field as well. The Linguistic String Project-Medical Language Processor is one the large scale projects of NLP in the field of medicine.

Approaches

Rationalist approach or symbolic approach assume that crucial part of the knowledge in the human mind is not derived by the sense but is firm in advance. It was trusted that machine can be made to function like human brain by giving some fundamental knowledge and reasoning mechanism linguistics knowledge is directly encoded in rule or other forms of representation.

Hidden Markov Model (HMM)

An HMM is a system where a shifting takes place between several states, generating feasible output symbols with each switch. The sets of viable states and unique symbols may be large, but finite and known. Pattern matching the state-switch sequence is realised are most likely to have generated a particular output-symbol sequence. Training the output-symbol chain data, reckon the state-switch/output probabilities that fit this data best. Hidden Markov Models are extensively used for speech recognition.The choice of area is wide ranging covering usual items like word segmentation and translation but also unusual areas like segmentation for infant learning and identifying documents for opinions and facts. In addition, exclusive article was selected for its use of Bayesian methods to aid the research in designing algorithms for their investigation.

5. NLP IN TALK

This section discusses the recent developments in the NLP projects implemented by various companies and these are as follows:

5.1 ACE Powered GDPR Robot Launched by RAVN Systems

RAVN Systems, an leading expert in Artificial Intelligence (AI), Search and Knowledge Management Solutions, announced the launch of a RAVN ("Applied Cognitive Engine") i.e powered software Robot to help and facilitate the GDPR ("General Data Protection Regulation") compliance.

5.2 Eno A Natural Language Chatbot Launched by Capital One

Capital one announces chatbot for customers called Eno. Eno is a natural language chatbot that people socialize through texting. Capital one claims that Eno is First natural language SMS chatbot from a U.S. bank that allows customer to ask questions using natural language. Customers can interact with Eno asking questions about their savings and others using a text interface. Eno makes such an environment that it feels that a human is interacting.

Future of BI in Natural Language Processing

BI will also make it easier to access as GUI is not needed. Because now a days the queries are made by text or voice command on smartphones.one of the most common example is Google might tell you today what will be the tomorrows weather. But soon enough, we will be able to ask our personal data (chatbot) about customer sentiment today, and how do we feel about their brand next week; all while walking down the street. Today, NLP tends to be based on turning natural language into machine language. But with time the technology matures – especially the AI component – the computer will get better at “understanding” the query and start to deliver answers rather than search results.

Meet the Pilot, world’s first language translating earbuds

The world’s first smart earpiece Pilot will soon be transcribed over 15 languages. According to Spring wise, Waverly Labs’ Pilot can already transliterate five spoken languages, English, French, Italian, Portuguese and Spanish, and seven written affixed languages, German, Hindi, Russian, Japanese, Arabic, Korean and Mandarin Chinese. The Pilot earpiece is connected via Bluetooth to the Pilot speech translation app, which uses speech recognition, machine translation and machine learning and speech synthesis technology.

REFERENCES

[1] Chomsky, Noam. Aspects of the Theory of Syntax.MASSACHUSETTS INST OF TECH CAMBRIDGE RESEARCH LAB OF ELECTRONICS, 1964.

[2] Rospocher, Marco, et al. "Building event-centric knowledge graphs from news." Web Semantics: Science, Services and Agents on the World Wide Web 37 (2016): 132-151.

[3] Shemtov, Hadar. Ambiguity management in natural language generation.Stanford University, 1997.

[4] Emele, Martin C., and Michael Dorna. "Ambiguity preserving machine translation using packed representations." Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational

Linguistics-Volume 1.Association for Computational Linguistics, 1998.

[5] Knight, Kevin, and Irene Langkilde. "Preserving ambiguities in generation via automata intersection."AAAI/IAAI. 2000.

[6] Nation, Kate, Margaret J. Snowling, and Paula Clarke. "Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension." Advances in Speech Language Pathology9.2 (2007): 131-139.

[7] Liddy, Elizabeth D. "Natural language processing." (2001).

[8] Feldman, Susan. "NLP meets the Jabberwocky: Natural language processing in information retrieval." ONLINE-WESTON THEN WILTON- 23 (1999): 62-73.

[9] "Natural Language Processing." Natural Language Processing RSS.N.p., n.d. Web. 25 Mar. 2017

[10] Hutchins, William John. Machine translation: past, present, future. Chichester: Ellis Horwood, 1986.

[11] Hutchins, W. John, ed. Early years in machine translation: memoirs and biographies of pioneers. Vol. 97.John Benjamins Publishing, 2000.

[12] Green Jr, Bert F., et al. "Baseball: an automatic question-answerer." Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. ACM, 1961.

[13] Woods, William A. "Semantics and quantification in natural language question answering." Advances in computers.Vol. 17.Elsevier, 1978.1-87.

[14] Hendrix, Gary G., et al. "Developing a natural language interface to complex data." ACM Transactions on Database Systems (TODS) 3.2 (1978): 105-147.

[15] Alshawi, Hiyan, ed. The core language engine. MIT press, 1992.

[16] Kamp, Hans, and UweReyle. "From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation." Studies in Linguistics and Philosophy.Kluwer (1993).

[17] Lea , W.A Trends in speech recognition , Englewoods Cliffs , NJ: Prentice Hall , 1980.

[18] Young, Steve J., and Lin Lawrence Chase. "Speech recognition evaluation: a review of the US CSR and LVCSR programmes." Computer Speech & Language 12.4 (1998): 263-279.

[19] Sundheim, Beth M., and Nancy A. Chinchor. "Survey of the message understanding conferences."Proceedings of the workshop on Human Language Technology.Association for Computational Linguistics, 1993.

[20] Wahlster, Wolfgang, and Alfred Kobsa. "User models in dialog systems." User models in dialog systems. Springer, Berlin, Heidelberg, 1989.4-34.

[21] McKeown, K.R. Text generation , Cambridge: Cambridge University Press , 1985.

[22] Small S.L., Cortell G.W., and Tanenhaus , M.K. Lexical Ambiguity Resolutions , San Mateo , CA : Morgan Kauffman, 1988.

[23] Manning, Christopher D., and HinrichSchütze. Foundations of statistical natural language processing. MIT press, 1999.

[24] Mani, Inderjeet, and Mark T. Maybury. Advances in automatic text summarization. MIT press, 1999.

[25] Yi, Jeonghee, et al. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques." Data Mining, 2003.ICDM 2003.Third IEEE International Conference on.IEEE, 2003.

[26] Yi, Jeonghee, et al. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques." Data Mining, 2003.ICDM 2003.Third IEEE International Conference on.IEEE, 2003.

[27] Tapaswi, Namrata, and Suresh Jain. "Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences." Software Engineering (CONSEG), 2012 CSI Sixth International Conference on. IEEE, 2012.

[28] Ranjan, Pradipta, and Harish V. SudeshnaSarkarAnupamBasu. "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi."Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003). 2003.

[29] Diab, Mona, KadriHacioglu, and Daniel Jurafsky. "Automatic tagging of Arabic text: From raw text to base phrase chunks." Proceedings of HLT-NAACL 2004: Short papers. Association for Computational Linguistics, 2004.

[30] Sha, Fei, and Fernando Pereira. "Shallow parsing with conditional random fields."Proceedings of the 2003 Conference of

the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.Association for Computational Linguistics, 2003.