NLP REVIEW PAPER

DEV PATEL – 21BCE050
CHINMAY TRIVEDI – 21BCE041

NATURAL LANGUAGE PROCESSING

—

2CS304

—

DIGITAL COMMUNICATION

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to all who directly or indirectly helped us to complete innovative assignment project under the digital communication course of third semester.

We, deeply express our sincerest thanks to

PROF. PARITA OZA

for her noble guidance.

Lastly thanking

Nirma University

for providing us with this opportunity.

THANK YOU

1. Abstract

Natural Language Processing (NLP) also known as "computational linguistics" is a type of artificial intelligence, which aims to deal with various procedures so that the computer is able to analyze, understand and generate natural human languages outputs seamlessly. The stream compiles the capabilities of semantic (interpreting words, signs etc.) & syntax (studying the phrasing of words so as to make a meaningful statement out of it). Furthermore, fields like audio recognition, automated translation, automatic summarization and many more fall under the applications of NLP. The review flips through the pages of history to analyses the evolvement of NLP and hence discusses modern aspects of the same. The primary objective that distinguishes the paper is usage of relatively simpler terminology. The paper aims to give a very apt & crisp idea about what, how and why NLP.

2. Introduction

Alan Turing – the father of computation once said, "A computer would deserve to be called intelligent if it could deceive a human into believing it was human." That is the very aim of Natural Language Processing. Researchers & Engineers have made great attempts in this field. They have combined the linguistic fields of semantics and syntax with powerful computer programs using neural network processes that "learn" how to look for the same kinds of signals humans look for to create meaning. This process of imitating a computer program to a humanly form is largely, single-handedly fueled by NLP. Research in NLP is highly inter-disciplinary. It encompasses concepts of Mathematics, linguistics, logic and psychology. NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

2.1 There are two main phases to NLP.

1) Data pre-processing – Data pre-processing involves preparing and cleaning the text data from the machines to be able to analyze it. It puts the data in a workable form so that an algorithm can work with it. This can be achieved via
    - Tokenization - This is when text is broken down into tokens (smaller units to work with.)
    - Stop word removal - This is when common words are removed from text so unique words that offer the most information about the text remain.
    - Lemmatization and stemming - This is when words are reduced to their root forms to process.
    - Part of speech tagging - This is when words are marked based on the part-of speech they are -- such as nouns, verbs and adjectives.

2) Once the data has been pre-processed, an algorithm is developed to process it. There are many different natural language processing algorithms, but two main types are commonly used:
    - Rule based system - This system uses carefully designed linguistic rules. This approach was used early on in the development of natural language processing, and is still used.
    - Machine learning based system - Machine learning algorithms use statistical methods. They learn to perform tasks based on training data they are fed, and adjust their methods as more data is processed. Using a combination of machine learning, deep learning and neural networks, natural language processing algorithms hone their own rules through repeated processing and learning.

3. Techniques

Syntax and semantic analysis are two main techniques used with natural language processing.

3.1 **Syntax**

Syntax is the arrangement of words in a sentence to make grammatical sense. NLP uses syntax to assess meaning from a language based on grammatical rules. Syntax techniques include:

- Parsing - This is the grammatical analysis of a sentence. Example: A natural language processing algorithm is fed the sentence, "The car is parked." Parsing involves breaking this sentence into parts of speech -- i.e., car = noun, parked = verb. This is useful for more complex downstream processing tasks.

- Word segmentation - This is the act of taking a string of text and deriving word forms from it. Example: A person scans a handwritten document into a computer. The algorithm would be able to analyze the page and recognize that the words are divided by white spaces.
- Sentence breaking - This places sentence boundaries in lacrge texts. Example: A natural language processing algorithm is fed the text, "The car is barked. It is in B2." The algorithm can recognize the period that splits up the sentences using sentence breaking.
- Morphological segmentation - This divides words into smaller parts called morphemes. Example: The word 'untestably' would be broken into [[un[[test]able]]ly], where the algorithm recognizes "un," "test," "able" and "ly" as morphemes. This is especially useful in machine translation and speech recognition. These mono syllable are used to detect the relevant words.
- Stemming - This divides words with inflection in them to root forms. Example: In the sentence, "The car is parked," the algorithm would be able to recognize the root of the word "parked" is "park." This would be useful if a user was analyzing a text for all instances of the word park, as well as all of its conjugations. The algorithm can see that they are essentially the same word even though the letters are different.

3.2 Semantics

Semantics involves the use of and meaning behind words. Natural language processing applies algorithms to understand the meaning and structure of sentences. Semantics techniques include:

- Word sense disambiguation - This derives the meaning of a word based on context. Example: Consider the sentence, "The pig is in the pen." The word pen has different meanings. An algorithm using this method can understand that the use of the word pen here refers to a fenced-in area, not a writing implement.
- Named entity recognition - This determines words that can be categorized into groups. Example: An algorithm using this method could analyze a news article and identify all mentions of a certain company or product. Using the semantics of the text, it would be able to differentiate between entities that are visually the same. For instance, in the sentence, "Daniel McDonald's son went to McDonald's and ordered a Happy Meal," the algorithm could recognize the two instances of "McDonald's" as two separate entities -- one a restaurant and one a person.

- Natural language generation - This uses a database to determine semantics behind words and generate new text. Example: An algorithm could automatically write a summary of findings from a business intelligence platform, mapping certain words and phrases to features of the data in the BI platform. Another example would be automatically generating news articles or tweets based on a certain body of text used for training.

4. History

a. **The first phase -** The evolution of NLP is currently an on-going process. The earliest known form of NLP is termed Machine translation. The objective was very specific, it simply translated data from one language to another. It was executed by IBM- Georgetown demonstration of 1954 and it translated Russian to English. A journal was also published regarding the same. The MT(machine translation) witnessed a breakthrough after the Teddington International conference on Machine translation in 1961. It successfully reported independent researched that were carried out by that time in different parts of the world. It included study in morphology, syntax and semantics. The main line of work during this period can be summarized as starting with translation as lookup, in dictionary-based word-for-word processing. The need to resolve syntactic and semantic ambiguity, and the former in particular because it is not open to fudging through the use of broad output equivalents, led to ambiguity resolution strategies based on local context, so dictionary entries became in effect individual procedures.

b. **The second phase -** The second phase of NLP was the AI powered, with the help of researches that took form of AI, the BASEBALL question answer game came into existence. The need to clarify the language user's objectives was early figured out by the Yale group, and since has become a major trend in NLP research since. Work on interactive dialogue in particular, from the second half of the 70s, has emphasised the communicative functionality, indirect function and underlying meaning, as well as direct function and surface meaning, of any linguistic expressions. At the same time work on discourse understanding in the 70s, whether on single-source texts like stories or reports, or on dialogue, maintenance and use of discourse models not relying only on prior scenarios like

scripts; and some useful progress was made with the development of notions of discourse or focus spaces and of resolution algorithms tied to these.

c. **The third phase -** The third phase of NLP development may be characterized as a grammatic-logical phase, whereas the second phase had an AI flavour and was semantics-focused. This movement was sparked by the growth of grammatical theory among linguists throughout the 1970s and the tendency towards using logic for knowledge representation and reasoning in artificial intelligence. It was a reaction to the shortcomings of practical system creation. Linguists produced a wide range of grammar types, such as functional, categorial, and generalised phrase structure, which, since they are focused on computability as an abstract goal, follow enhanced transition networks as computational grammars in both a theoretical and practical sense. The work on the lexicon increased significantly, which was the last 80s trend. The necessity for multilingual MT, the significant role the lexicon plays in the grammatic-logical approach, as well as the issues with portability, customization, and knowledge acquisition in connection to specific applications, all served as catalysts for this. The first significant attempts to use commercial dictionaries in machine-readable form have now been made, and this has ultimately led to the use of text corpora to validate, enhance, or customise initial lexical data. Research has been made significantly easier by the rapidly expanding supply of text material. Now, it can be observed that this last tendency is what gives the present fourth period of NLP its primary color.

d. **Current** (expanded in 5th subheading) - Finally, the fourth period has seen new interest in multi-modal, or multimedia, systems. This is in part a natural response to the opportunities offered by modem computing technology, and in part an attempt to satisfy human needs and skills in information management. But whether combining language with other modes or media, like graphics, actually simplifies or complicates language processing is an open question.

5. *Where are we now*

With the evolution of NLP, speech recognition systems are using deep neural networks. Different vowels or sounds have different frequencies, which are discernible on a spectrogram.

This allows computers to recognize spoken vowels and words. Each sound is called a phoneme, and speech recognition software knows what these phonemes look like. Along with analyzing different words, NLP helps discern where sentences begin and end. And ultimately, speech is converted to text. Speech synthesis gives computers the ability to output speech. However, these sounds are discontinuous and seem robotic. While this was very prominent in the hand-operated machine from Bell Labs, today's computer voices like Siri and Alexa have improved.

We are now seeing an explosion of voice interfaces on phones and cars. This creates a positive feedback loop with people using voice interaction more often, which gives companies more data to work on. This enables better accuracy, leading to people using voice more, and the loop continues. NLP evolution has happened by leaps and bounds in the last decade. NLP integrated with deep learning and machine learning has enabled chatbots and virtual assistants to carry out complicated interactions. Chatbots now operate beyond the domain of customer interactions. They can handle human resources and healthcare, too. NLP in healthcare can monitor treatments and analyze reports and health records. Cognitive analytics and NLP are combined to automate routine tasks.

Current approaches to natural language processing are based on deep learning, a type of AI that examines and uses patterns in data to improve a program's understanding. Deep learning models require massive amounts of labeled data for the natural language processing algorithm to train on and identify relevant correlations, and assembling this kind of big data set is one of the main hurdles to natural language processing.

Earlier approaches to natural language processing involved a more rules-based approach, where simpler machine learning algorithms were told what words and phrases to look for in text and given specific responses when those phrases appeared. But deep learning is a more flexible, intuitive approach in which algorithms learn to identify speakers' intent from many examples -- almost like how a child would learn human language.

Three tools used commonly for natural language processing include Natural Language Toolkit (NLTK), Gensim and Intel natural language processing Architect. NLTK is an open source Python module with data sets and tutorials. Gensim is a Python library for topic modeling and

document indexing. Intel NLP Architect is another Python library for deep learning topologies and techniques.

6. Basic Algorithms

The evolution of NLP has happened with time and advancements in language technology. Data scientists developed some powerful algorithms along the way; some of them are as follows:

- Bag of words: This model counts the frequency of each unique word in an article. This is done to train machines to understand the similarity of words. However, millions of individual words are in millions of documents; hence, maintaining such vast data is practically unimaginable.
  //sample code for demonstration

```
def vectorize(tokens):
    ''' This function takes list of words in a sentence as input
    and returns a vector of size of filtered_vocab.It puts 0 if the
    word is not present in tokens and count of token if present.'''
    vector=[]
    for w in filtered_vocab:
        vector.append(tokens.count(w))
    return vector
def unique(sequence):
    '''This functions returns a list in which the order remains
    same and no item repeats.Using the set() function does not
    preserve the original ordering,so i didnt use that instead'''
    seen = set()
    return [x for x in sequence if not (x in seen or seen.add(x))]
#create a list of stopwords.You can import stopwords from nltk too
stopwords=["to","is","a"]
#list of special characters.You can use regular expressions too
special_char=[",",":"," ",";",".","?"]
#Write the sentences in the corpus,in our case, just two
string1="We are going to Nirma and today is a nice day"
string2="Parita ma'am is our faculty for DC"
#convert them to lower case
string1=string1.lower()
string2=string2.lower()
#split the sentences into tokens
tokens1=string1.split()
```

```
tokens2=string2.split()
print(tokens1)
print(tokens2)
#create a vocabulary list
vocab=unique(tokens1+tokens2)
print(vocab)
#filter the vocabulary list
filtered_vocab=[]
for w in vocab:
    if w not in stopwords and w not in special_char:
        filtered_vocab.append(w)
print(filtered_vocab)
#convert sentences into vectords
vector1=vectorize(tokens1)
print(vector1)
vector2=vectorize(tokens2)
print(vector2)
```

- TF-IDF: TF (term frequency) is calculated as the number of times a certain term appears out of the number of terms present in the document. This system also eliminates "stop words," like "is," "a," "the," etc.
  Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency:
  Number of times the word appears in a document (raw count).
  Term frequency adjusted for the length of the document (raw count of occurences divided by number of words in the document).
  Logarithmically scaled frequency (e.g. log(1 + raw count)).
  Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

$$idf\,(t, D) = log\,(\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

  IDF is the inverse document frequency.

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Here the sample statement was
"The car is driven on the road" and "The truck is driven on the highway".

- Co-occurrence matrix: This model was developed since the previous models could not solve the problem of semantic ambiguity. It tracked the context of the text but required a lot of memory to store all the data.

  For example the statement is "*Roses are red. Sky is blue"*

	Roses	are	red	Sky	is	blue
Roses	1	1	1	0	0	0
are	1	1	1	0	0	0
red	1	1	1	0	0	0
Sky	0	0	0	1	1	1
is	0	0	0	1	1	1
Blue	0	0	0	1	1	1

- Transformer models: This is the encoder and decoder model that uses attention to train the machines that imitate human attention faster. BERT, developed by Google based on this model, has been phenomenal in revolutionizing NLP.

  Here is an example to demonstrate the significance of transformer model.
  "I poured water from the bottle into the cup until it was full."
  it => cup
  "I poured water from the bottle into the cup until it was empty."
  it=> bottle

7. Importance of NLP

NLP solves the root problem of machines not understanding human language. With its evolution, NLP has surpassed traditional applications, and AI is being used to replace human resources in several domains.

- Machine translation" is a significant application of NLP. NLP is behind the widely used Google Translate, which converts one language into another in real-time. It assists computers in understanding the context of sentences and the meaning of words.
- Virtual assistants like Cortana, Siri, and Alexa are boons of NLP evolution. These assistants comprehend what you say, give befitting replies, or take appropriate actions, and do all this through NLP.
- Intelligent chatbots are taking the world of customer service by storm. They are replacing human assistance and conversing with customers just like humans do. They interpret the written text, and it decides on actions accordingly. NLP is the working mechanism behind such chatbots.
- NLP also helps in sentiment analysis. It recognizes the sentiment behind posts. For instance, it determines whether a review is positive, negative, serious or sarcastic. NLP mechanisms help companies like Twitter remove tweets with foul language, etc.
- NLP automatically sorts our emails into social, promotions, inbox, and spam categories. This NLP task is known as text classification.
- Other applications of NLP are seen in checking spellings, keyword research, and extracting information. Plagiarism checkers also run-on NLP programs.
- NLP also drives advertisement recommendations. It matches advertisements with our history.
- NLP helps machines understand natural languages and perform language-related tasks. It makes it possible for computers to analyse more language-based data than humans.

8. Challenges

Challenges of natural language processing
There are a number of challenges of natural language processing and most of them boil down to the fact that natural language is ever-evolving and always somewhat ambiguous. They include:

Precision.
Computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured -- or through a limited number of clearly

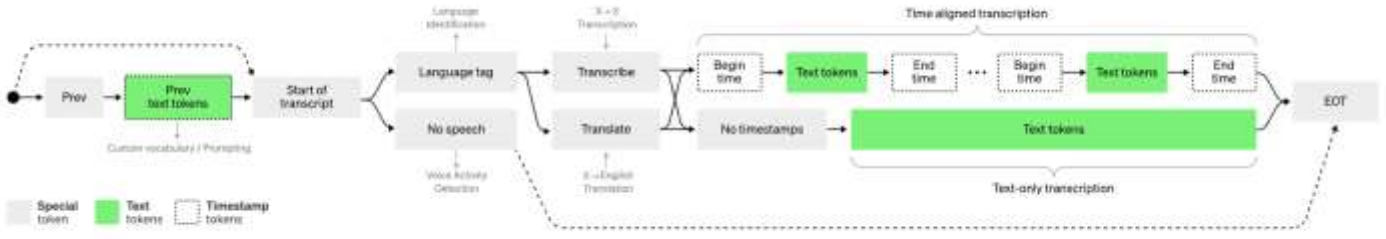
enunciated voice commands. Human speech, however, is not always precise; it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

Tone of voice and inflection.
Natural language processing has not yet been perfected. For example, semantic analysis can still be a challenge. Other difficulties include the fact that the abstract use of language is typically tricky for programs to understand. For instance, natural language processing does not pick up sarcasm easily. These topics usually require understanding the words being used and their context in a conversation. As another example, a sentence can change meaning depending on which word or syllable the speaker puts stress on. NLP algorithms may miss the subtle, but important, tone changes in a person's voice when performing speech recognition. The tone and inflection of speech may also vary between different accents, which can be challenging for an algorithm to parse.

Evolving use of language.
Natural language processing is also challenged by the fact that language -- and the way people use it -- is continually changing. Although there are rules to language, none are written in stone, and they are subject to change over time. Hard computational rules that work now may become obsolete as the characteristics of real-world language change over time.

9. Special application

Whisper is an trained open AI that approaches human level robustness and accuracy on English speech recognition. It is an automated speech recognition (ASR) system that has been trained on 6,8o,ooo hours of supervised, multilingual, and multitask online data. Whisper demonstrates how using a data set this size and variety increases robustness against accents, background noise and technical terminology. Additionally, it permits both translation into English from several languages as well as transcription in those languages.



https://openai.com/blog/whisper

10. Future prospects

We may anticipate substantially more robust spoken language systems that also contain certain characteristics of discourse structure as the trend toward speech and language integration continues. In the near future, more sophisticated and practical commercial systems for machine translation, message interpretation, and multilingual interfaces will emerge. Systems for translating voice to speech will also start to show up in some niches. Language systems will be beneficial for various applications, including giving directions for the assembling or maintaining sophisticated machinery, and will become increasingly more integrated with other modalities like visuals (including pointing). Additionally, NLP systems will be crucial components of educational and prosthetic systems. On the theoretical front, research in mathematics and computation has provided us with rich insights into how language functions. But the structure of language is incredibly intricate. The computational knowledge of the structure and function of language is thus still in a very early stage. We will have more comprehensive explanations of the syntax, semantics, and pragmatic aspects of language as a result of more mathematical and computational studies. We will see a further increase in the use of language corpora and statistical approaches in NLP, as well as a merger of structural and statistical techniques in NLP that will result in more reliable systems. Finally, research in psycholinguistics, which examines how people perceive language, will benefit from the mathematical, computational, and statistical work done in NLP. This work might lead to a breakthrough in NLP.

11. References

[1] Karen Sparck Jones, Natural Language Processing – A historical review, university of Cambridge
[2] A1shawi, H. (ed), The Core Language Engine, Cambridge, MA: MIT Press, 1992.
[3] Bledsoe, W. "I had a dream: AAAI presidential address, 19 August 1985", The AI Magazine 7 (1), 1986, 57-61.
[4] Bobrow, D.G. and Collins, A. (eds) Representation and understanding, New York: Academic, 1975.
[5] Booth, A.D. (ed.) Machine translation, Amsterdam: North-Holland, 1967.
[6] Brady, M. and Berwick, R.C. (eds.) Computational models of discourse, Cambridge, MA: MIT Press, 1983.
[7] Briscoe, E. et al. "A formalism and environment for the development of a large grammar of English", IJCAI 87: Proceedings of the lOth International Joint Conference on Artificial Intelligence, 1987, 703-708.
[8] Ceccato, S. "Correlational analysis and mechanical translation", in Booth 1967, 77-135.
[9] Cohen, P.R., Morgan, J. and Pollack, M.E. (eds.) Intentions in communication, Cambridge, MA: MIT Press, 1990.
[10] Cullingford, R. "SAM", 1981; reprinted in Grosz et al. 1986, 627-649.
[11] Engelien, B. and McBryde, R. Natura/language markets: commercial strategies, Ovum Ltd, 7 Rathbone Street, London, 1991.
[12] Findler, N.V. (ed.) Associative networks, New York: Academic, 1979.

[13] Galliers, J.R. and Sparck Jones, K. Evaluating natura/language processing systems, Technical Report 291, Computer Laboratory, University of Cambridge, 1993.
[14] Green, B.F. et al "BASEBALL: an automatic question answerer", 1961; reprinted in Grosz et al., 1986, 545-549.
[15] Grosz, B.J., Sparck Jones, K. and Webber, B.L. (eds) Readings in natural language processing, Los Altos, CA: Morgan Kaufmann, 1986.
[16] Harris, L.R. "Experience with INTELLECT', The AI Magazine 5(2), 1984, 43-50. [17] Hays, D.G. Introduction to computational linguistics, London: Macdonald, 1967.
[18] Hendrix, G., Sacerdoti, E., Sagalowicz, D., Slocum, J., "Developing a Natural Language Interface to Complex Data", ACM Transactions on Database Systems,
Vol3, No.3, pp 105-147, 1978.
HLT: Proceedings of the ARPA Workshop on Human Language Technology, March 1993; San Mateo, CA: Morgan Kaufmann, in press.
[20] Hutchins, W.J. Machine translation, Chichester, England: Ellis Horwood, 1986.
[21] Hutchins, W.J. and Somers, H.L. An introduction to machine translation, London: Academic Press, 1992.
[22] Jacobs, P.S. (ed) Text-based intelligent systems, Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.
[23] Joshi, A.K., Webber, B.L. and Sag, I.A. (eds.) Elements of discourse understanding, Cambridge: Cambridge University Press, 1981.
[24] Kay, M., Gawron, J.M. and Norvig, P. Verbmobil: a translation system for faceto-face dialogue, CSLI, Stanford University, 1991.
[25] Kittredge, R. and Lehrberger, J. (eds.) Sublanguage: studies of language in restricted semantic domains, Berlin; Walter de Gruyter, 1982.
[26] Kobsa, A. and Wahlster, W. (eds.) User modelling in dialogue systems, Berlin: Springer-Verlag, 1989.
[27] Lea, W.A. (ed) Trends in speech recognition, Englewood Cliffs, NJ: Prentice-Hall, 1980.
[28] Locke, W.N. and Booth, A.D. (eds.) Machine translation of languages, New York: John Wiley, 1955.
[29] McKeown, K.R. Text generation, Cambridge: Cambridge University Press, 1985.
[30] Minsky, M. (ed.) Semantic information processing, Cambridge, MA: MIT Press, 1968.
[31] Minsky, M., "A framework for representing knowledge," (ed Winston, P.), The psychology of computer vision, McGraw-Hill, 1975.
[32] Nagao, M. (ed) A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, USA, Japan Electronic Industry Development Association, 1989.
[33] Plath, W. "Multiple path analysis and automatic translation", in Booth 1967,267315.
[34] Reifler, E. "Chinese-English machine translation, its lexicographic and linguistic problems", in Booth 1967,317-428.
[35] Rumelhart, D.E., McClelland, J.L. and the PDP Research Group, Parallel distributed processing, 2 vols, Cambridge, MA: MIT Press, 1986.
[36] Rustin, R. (ed) Natura/language processing, New York: Algorithmics Press, 1973.
[37] The American Archivist, Vol. 61, No. 2 (Fall, 1998), pp. 400-425, Jane Greenberg, The applicability of NLP, Fall – 1998.