Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13428-4

Natural language processing: state of the art, current trends and challenges

Diksha Khurana[1] · Aditya Koli[1] · Kiran Khatter[2] · Sukhdev Singh[3]

Received: 3 February 2021 / Revised: 23 March 2022 / Accepted: 2 July 2022

Abstract

Natural language processing (NLP) has recently gained much attention for representing and analyzing 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. In this paper, we first distinguish four phases by discussing different levels of NLP and components of **N**atural **L**anguage **G**eneration followed by presenting the history and evolution of NLP. We then discuss in detail the state of the art presenting the various applications of NLP, current trends, and challenges. Finally, we present a discussion on some available datasets, models, and evaluation metrics in NLP.

Keywords Natural language processing · Natural language understanding · Natural language generation · NLP applications · NLP evaluation metrics

✉ Kiran Khatter
kirankhatter@gmail.com

Diksha Khurana
dikshakhurana0509@gmail.com

Aditya Koli
adityakoli0010@gmail.com

Sukhdev Singh
sukhdev200@gmail.com

[1] Department of Computer Science, Manav Rachna International Institute of Research and Studies, Faridabad, India

[2] Department of Computer Science, BML Munjal University, Gurgaon, India

[3] Department of Statistics, Amity University Punjab, Mohali, India

Published online: 14 July 2022

1 Introduction

A language can be defined as a set of rules or set of symbols where symbols are combined and used for conveying information or broadcasting the information. Since all the users may not be well-versed in machine specific language, **N**atural Language Processing (NLP) caters those users who do not have enough time to learn new languages or get perfection in it. In fact, NLP is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages. It came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language, and can be classified into two parts i.e. *Natural Language Understanding* or Linguistics and *Natural Language Generation* which evolves the task to understand and generate the text. **L**inguistics is the science of language which includes *Phonology* that refers to sound, *Morphology* word formation, *Syntax* sentence structure, *Semantics* syntax and *Pragmatics* which refers to understanding. Noah Chomsky, one of the first linguists of twelfth century that started syntactic theories, marked a unique position in the field of theoretical linguistics because he revolutionized the area of syntax (Chomsky, 1965) [23]. Further, Natural Language Generation (NLG) is the process of producing phrases, sentences and paragraphs that are meaningful from an internal representation. The first objective of this paper is to give insights of the various important terminologies of NLP and NLG.

In the existing literature, most of the work in NLP is conducted by computer scientists while various other professionals have also shown interest such as linguistics, psychologists, and philosophers etc. One of the most interesting aspects of NLP is that it adds up to the knowledge of human language. The field of NLP is related with different theories and techniques that deal with the problem of natural language of communicating with the computers. Few of the researched tasks of NLP are Automatic Summarization (*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 (*Co-reference resolution* refers to a sentence or larger set of text that determines all words which refer to the same object), Discourse Analysis (*Discourse analysis* refers to the task of identifying the discourse structure of connected text i.e. the study of text in relation to social context),Machine Translation (*Machine translation* refers to automatic translation of text from one language to another),Morphological Segmentation (*Morphological segmentation* refers to breaking words into individual meaning-bearing morphemes), Named Entity Recognition (*Named entity recognition* (NER) is used for information extraction to recognized name entities and then classify them to different classes), Optical Character Recognition (*Optical character recognition* (OCR) is used for automatic text recognition by translating printed and handwritten text into machine-readable format), Part Of Speech Tagging (*Part of speech tagging* describes a sentence, determines the part of speech for each word) etc. Some of these tasks have direct real-world applications such as Machine translation, Named entity recognition, Optical character recognition etc. Though NLP tasks are obviously very closely interwoven but they are used frequently, for convenience. Some of the tasks such as automatic summarization, co-reference analysis etc. act as subtasks that are used in solving larger tasks. Nowadays NLP is in the talks because of various applications and recent developments although in the late 1940s the term wasn't even in existence. So, it will be interesting to know about the history of NLP, the progress so far has been made and some of the ongoing projects by making use of NLP. The second objective of this paper focus on these aspects. The third objective of this paper is on datasets, approaches, evaluation metrics and involved challenges in NLP. The rest of this

paper is organized as follows. Section 2 deals with the first objective mentioning the various important terminologies of NLP and NLG. Section 3 deals with the history of NLP, applications of NLP and a walkthrough of the recent developments. Datasets used in NLP and various approaches are presented in Section 4, and Section 5 is written on evaluation metrics and challenges involved in NLP. Finally, a conclusion is presented in Section 6.

2 Components of NLP

NLP 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. Figure 1 presents the broad classification of NLP. The objective of this section is to discuss the Natural Language Understanding (Linguistic) (NLU) and the Natural Language Generation **(NLG)**.

2.1 NLU

NLU enables machines to understand natural language and analyze it by extracting concepts, entities, emotion, keywords etc. It is used in customer care applications to understand the problems reported by customers either verbally or in writing. Linguistics is the science which involves the meaning of language, language context and various forms of the language. So, it is important to understand various important terminologies of NLP and different levels of NLP. We next discuss some of the commonly used terminologies in different levels of NLP.

a) Phonology

Phonology is the part of Linguistics which refers to the systematic arrangement of sound. The term phonology comes from Ancient Greek in which the term phono means voice or sound and the suffix –logy refers to word or speech. In 1993 Nikolai Trubetzkoy stated that

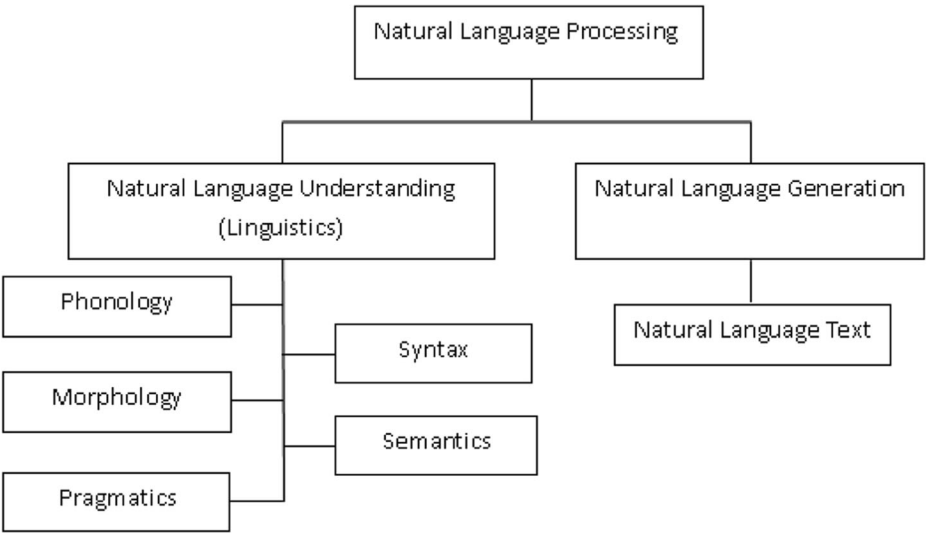

Fig. 1 Broad classification of NLP

Phonology is "the study of sound pertaining to the system of language" whereas Lass1998 [66]wrote that phonology refers broadly with the sounds of language, concerned with sub-discipline of linguistics, behavior and organization of sounds. Phonology includes semantic use of sound to encode meaning of any Human language.

b) Morphology

The different parts of the word represent the smallest units of meaning known as Morphemes. Morphology which comprises Nature of words, are initiated by morphemes. An example of Morpheme could be, the word *precancellation* can be morphologically scrutinized into three separate morphemes: the prefix *pre*, the root *cancella*, and the suffix *-tion*. The interpretation of morphemes stays the same across all the words, just to understand the meaning humans can break any unknown word into morphemes. For example, adding the suffix *–ed* to a verb, conveys that the action of the verb took place in the past. The words that cannot be divided and have meaning by themselves are called Lexical morpheme (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). The Grammatical morphemes that occur in combination called bound morphemes (eg. -ed, −ing) Bound morphemes can be divided into inflectional morphemes and derivational morphemes. Adding Inflectional morphemes to a word changes the different grammatical categories such as tense, gender, person, mood, aspect, definiteness and animacy. For example, addition of inflectional morphemes *–ed* changes the root *park* to *parked*. Derivational morphemes change the semantic meaning of the word when it is combined with that word. For example, in the word normalize, the addition of the bound morpheme *–ize* to the root normal changes the word from an adjective (*normal*) to a verb (*normalize*).

c) Lexical

In Lexical, humans, as well as NLP systems, interpret the meaning of individual words. Sundry types of processing bestow to word-level understanding – the first of these being a part-of-speech tag to each word. In this processing, words that can act as more than one part-of-speech are assigned the most probable part-of-speech tag based on the context in which they occur. At the lexical level, Semantic representations can be replaced by the words that have one meaning. In fact, in the NLP system the nature of the representation varies according to the semantic theory deployed. Therefore, at lexical level, analysis of structure of words is performed with respect to their lexical meaning and PoS. In this analysis, text is divided into paragraphs, sentences, and words. Words that can be associated with more than one PoS are aligned with the most likely PoS tag based on the context in which they occur. At lexical level, semantic representation can also be replaced by assigning the correct POS tag which improves the understanding of the intended meaning of a sentence. It is used for cleaning and feature extraction using various techniques such as removal of stop words, stemming, lemmatization etc. Stop words such as '*in*', 'the', 'and' etc. are removed as they don't contribute to any meaningful interpretation and their frequency is also high which may affect the computation time. Stemming is used to stem the words of the text by removing the suffix of a word to obtain its root form. For example: *consulting* and *consultant* words are converted to the word *consult* after stemming, *using* word gets converted to *us* and *driver* is reduced to *driv*. Lemmatization does not remove the suffix of a word; in fact, it results in the source word

with the use of a vocabulary. For example, in case of token *drived*, stemming results in "driv", whereas lemmatization attempts to return the correct basic form either *drive* or *drived* depending on the context it is used.

d) Syntactic

After PoS tagging done at lexical level, words are grouped to phrases and phrases are grouped to form clauses and then phrases are combined to sentences at syntactic level. It emphasizes the correct formation of a sentence by analyzing the grammatical structure of the sentence. The output of this level is a sentence that reveals structural dependency between words. It is also known as *parsing* which uncovers the phrases that convey more meaning in comparison to the meaning of individual words. Syntactic level examines word order, stop-words, morphology and PoS of words which lexical level does not consider. Changing word order will change the dependency among words and may also affect the comprehension of sentences. For example, in the sentences "ram beats shyam in a competition" and "shyam beats ram in a competition", only syntax is different but convey different meanings [139]. It retains the stopwords as removal of them changes the meaning of the sentence. It doesn't support lemmatization and stemming because converting words to its basic form changes the grammar of the sentence. It focuses on identification on correct PoS of sentences. For example: in the sentence "frowns on his face", "frowns" is a noun whereas it is a verb in the sentence "he frowns".

e) Semantic

On a semantic level, the most important task is to determine the proper meaning of a sentence. To understand the meaning of a sentence, human beings rely on the knowledge about language and the concepts present in that sentence, but machines can't count on these techniques. Semantic processing determines the possible meanings of a sentence by processing its logical structure to recognize the most relevant words to understand the interactions among words or different concepts in the sentence. For example, it understands that a sentence is about "movies" even if it doesn't comprise actual words, but it contains related concepts such as "actor", "actress", "dialogue" or "script". This level of processing also incorporates the semantic disambiguation of words with multiple senses (Elizabeth D. Liddy, 2001) [68]. For example, the word "bark" as a noun can mean either as a sound that a dog makes or outer covering of the tree. The semantic level examines words for their dictionary interpretation or interpretation is derived from the context of the sentence. For example: the sentence "Krishna is good and noble." This sentence is either talking about Lord Krishna or about a person "Krishna". That is why, to get the proper meaning of the sentence, the appropriate interpretation is considered by looking at the rest of the sentence [44].

f) Discourse

While syntax and semantics level deal with sentence-length units, the discourse level of NLP deals with more than one sentence. It deals with the analysis of logical structure by making connections among words and sentences that ensure its coherence. It focuses on the properties of the text that convey meaning by interpreting the relations between sentences and uncovering linguistic structures from texts at several levels (Liddy,2001) [68]. The two of the most common levels are: *Anaphora Resolution* and *Coreference Resolution*. Anaphora resolution

is achieved by recognizing the entity referenced by an anaphor to resolve the references used within the text with the same sense. For example, (i) Ram topped in the class. (ii) He was intelligent. Here i) and ii) together form a discourse. Human beings can quickly understand that the pronoun "he" in (ii) refers to "Ram" in (i). The interpretation of "He" depends on another word "Ram" presented earlier in the text. Without determining the relationship between these two structures, it would not be possible to decide why Ram topped the class and who was intelligent. Coreference resolution is achieved by finding all expressions that refer to the same entity in a text. It is an important step in various NLP applications that involve high-level NLP tasks such as document summarization, information extraction etc. In fact, anaphora is encoded through one of the processes called co-reference.

g) Pragmatic

Pragmatic level focuses on the knowledge or content that comes from the outside the content of the document. It deals with what speaker implies and what listener infers. In fact, it analyzes the sentences that are not directly spoken. Real-world knowledge is used to understand what is being talked about in the text. By analyzing the context, meaningful representation of the text is derived. When a sentence is not specific and the context does not provide any specific information about that sentence, Pragmatic ambiguity arises (Walton, 1996) [143]. Pragmatic ambiguity occurs when different persons derive different interpretations of the text, depending on the context of the text. The context of a text may include the references of other sentences of the same document, which influence the understanding of the text and the background knowledge of the reader or speaker, which gives a meaning to the concepts expressed in that text. Semantic analysis focuses on *literal meaning* of the words, but pragmatic analysis focuses on the *inferred meaning* that the readers perceive based on their background knowledge. For example, the sentence "Do you know what time is it?" is interpreted to "Asking for the current time" in semantic analysis whereas in pragmatic analysis, the same sentence may refer to "expressing resentment to someone who missed the due time" in pragmatic analysis. Thus, semantic analysis is the study of the relationship between various linguistic utterances and their meanings, but pragmatic analysis is the study of context which influences our understanding of linguistic expressions. Pragmatic analysis helps users to uncover the intended meaning of the text by applying contextual background knowledge.

The goal of NLP is to accommodate one or more specialties of an algorithm or system. The metric of NLP assess on an algorithmic system allows for the integration of language understanding and language generation. It is even used in multilingual event detection. Rospocher et al. [112] purposed a novel modular system 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 NLP tools. The pipeline integrates modules for basic NLP processing as well as more advanced tasks such as cross-lingual named entity linking, semantic role labeling and time normalization. Thus, the cross-lingual framework allows for the interpretation of events, participants, locations, and time, as well as the relations between them. Output of these individual pipelines is intended to be used as input for a system that obtains event centric knowledge graphs. All modules take standard input, to do some annotation, and produce standard output which in turn becomes the input for the next module pipelines. Their pipelines are built as a data centric architecture so that modules can be adapted and replaced. Furthermore, modular architecture allows for different configurations and for dynamic distribution.

Ambiguity is one of the major problems of natural language which occurs when one sentence can lead to different interpretations. This is usually faced in syntactic, semantic, and lexical levels. In case of syntactic level ambiguity, one sentence can be parsed into multiple syntactical forms. Semantic ambiguity occurs when the meaning of words can be misinterpreted. Lexical level ambiguity refers to ambiguity of a single word that can have multiple assertions. Each of these levels can produce ambiguities that can be solved by the knowledge of the complete sentence. The ambiguity can be solved by various methods such as Minimizing Ambiguity, Preserving Ambiguity, Interactive Disambiguation and Weighting Ambiguity [125]. Some of the methods proposed by researchers to remove ambiguity is preserving ambiguity, e.g. (Shemtov 1997; Emele & Dorna 1998; Knight & Langkilde 2000; Tong Gao et al. 2015, Umber & Bajwa 2011) [39, 46, 65, 125, 139]. Their objectives are closely in line with removal or minimizing ambiguity. They cover a wide range of ambiguities and there is a statistical element implicit in their approach.

2.2 NLG

Natural Language Generation (NLG) is the process of producing phrases, sentences and paragraphs that are meaningful from an internal representation. It is a part of Natural Language Processing and happens in four phases: identifying the goals, planning on how goals may be achieved by evaluating the situation and available communicative sources and realizing the plans as a text (Fig. 2). It is opposite to Understanding.

a) Speaker and Generator

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

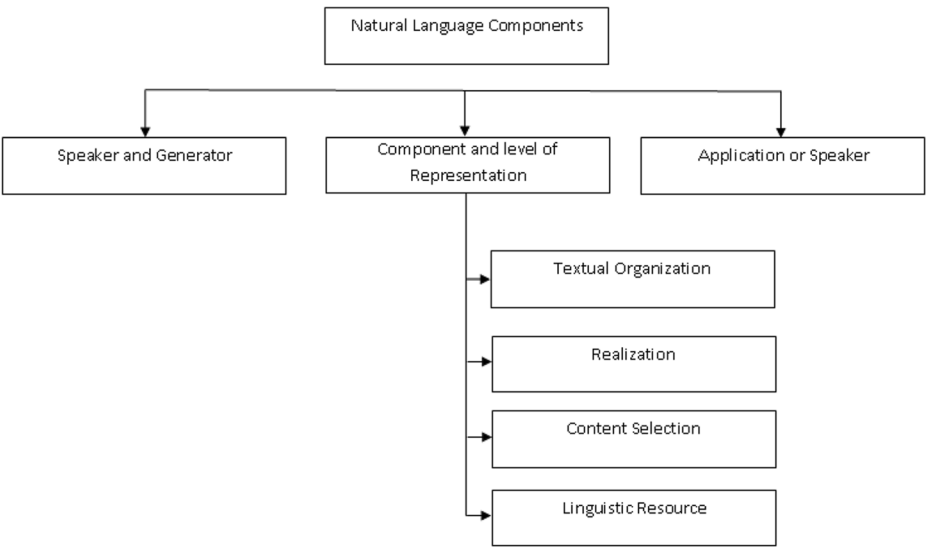

Fig. 2 Components of NLG

b) Components and Levels of Representation

The process of language generation involves the following interweaved tasks. *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 to 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.

c) Application or Speaker

This is only for maintaining the model of the situation. Here the speaker just initiates the process doesn't take part in the language generation. It stores the history, structures the content that is potentially relevant and deploys a representation of what it knows. All these forms the situation, while selecting subset of propositions that speaker has. The only requirement is the speaker must make sense of the situation [91].

3 NLP: Then and now

In the late 1940s the term NLP wasn't in existence, but the work regarding machine translation (MT) had started. In fact, Research in this period was not completely localized. Russian and English were the dominant languages for MT (Andreev,1967) [4]. In fact, MT/NLP research almost died in 1966 according to the ALPAC report, which concluded that MT is going nowhere. But later, some MT production systems were providing output to their customers (Hutchins, 1986) [60]. By this time, work on the use of computers for literary and linguistic studies had also started. As early as 1960, signature work influenced by AI began, with the BASEBALL Q-A systems (Green et al., 1961) [51]. LUNAR (Woods,1978) [152] and Winograd SHRDLU were natural successors of these systems, but they were seen as stepped-up sophistication, in terms of their linguistic and their task processing capabilities. There was a widespread belief that progress could only be made on the two sides, one is ARPA Speech Understanding Research (SUR) project (Lea, 1980) and other in some major system developments projects building database front ends. The front-end projects (Hendrix et al., 1978) [55] were intended to go beyond LUNAR in interfacing the large databases. In early 1980s computational grammar theory became a very active area of research linked with logics for meaning and knowledge's ability to deal with the user's beliefs and intentions and with functions like emphasis and themes.

By the end of the decade the powerful general purpose sentence processors like SRI's Core Language Engine (Alshawi,1992) [2] and Discourse Representation Theory (Kamp and Reyle,1993) [62] offered a means of tackling more extended discourse within the grammatico-logical framework. This period was one of the growing communities. Practical resources, grammars, and tools and parsers became available (for example: Alvey Natural Language Tools) (Briscoe et al., 1987) [18]. The (D)ARPA speech recognition and message understanding (information extraction) conferences were not only for the tasks they addressed

but for the emphasis on heavy evaluation, starting a trend that became a major feature in 1990s (Young and Chase, 1998; Sundheim and Chinchor,1993) [131, 157]. Work on user modeling (Wahlster and Kobsa, 1989) [142] was one strand in a research paper. Cohen et al. (2002) [28] had put forwarded a first approximation of a compositional theory of tune interpretation, together with phonological assumptions on which it is based and the evidence from which they have drawn their proposals. At the same time, McKeown (1985) [85] demonstrated that rhetorical schemas could be used for producing both linguistically coherent and communicatively effective text. Some research in NLP marked important topics for future like word sense disambiguation (Small et al., 1988) [126] and probabilistic networks, statistically colored NLP, the work on the lexicon, also pointed in this direction. Statistical language processing was a major thing in 90s (Manning and Schuetze,1999) [75], because this not only involves data analysts. Information extraction and automatic summarizing (Mani and Maybury,1999) [74] was also a point of focus. Next, we present a walkthrough of the developments from the early 2000.

3.1 A walkthrough of recent developments in NLP

The main objectives of NLP include interpretation, analysis, and manipulation of natural language data for the intended purpose with the use of various algorithms, tools, and methods. However, there are many challenges involved which may depend upon the natural language data under consideration, and so makes it difficult to achieve all the objectives with a single approach. Therefore, the development of different tools and methods in the field of NLP and relevant areas of studies have received much attention from several researchers in the recent past. The developments can be seen in the Fig. 3:

In early 2000, *neural language modeling* in which the probability of occurring of next word (token) is determined given *n* previous words. Bendigo et al. [12] proposed the concept of feed forward neural network and lookup table which represents the n previous words in sequence. Collobert et al. [29] proposed the application of *multitask learning* in the field of NLP, where two convolutional models with max pooling were used to perform parts-of-speech and named entity recognition tagging. Mikolov et.al. [87] proposed a *word embedding* process where the dense vector representation of text was addressed. They also report the challenges faced by traditional sparse bag-of-words representation. After the advancement of word embedding, neural networks were introduced in the field of NLP where variable length input is taken for further processing. Sutskever et al. [132] proposed a general framework for *sequence-to-sequence* mapping where encoder and decoder networks are used to map from sequence to vector and vector to sequence respectively. In fact, the use of *neural networks* have played a very important role in NLP. One can observe from the existing literature that enough use of neural networks was not there in the early 2000s but till the year 2013enough discussion had happened about the use of neural networks in the field of NLP which transformed many things and further paved the way to implement various neural networks in NLP. Earlier the use of *Convolutional neural networks* **(**CNN**)** contributed to the field of image classification and analyzing visual imagery for further analysis. Later the use of CNNs can be observed in tackling problems associated with NLP tasks like Sentence Classification [127], Sentiment Analysis [135], Text Classification [118], Text Summarization [158], Machine Translation [70] and Answer Relations [150]**.** An article by Newatia (2019) [93] illustrates the general architecture behind any CNN model, and how it can be used in the context of NLP. One can also refer to the work of Wang and Gang [145] for the applications of CNN in NLP. Further

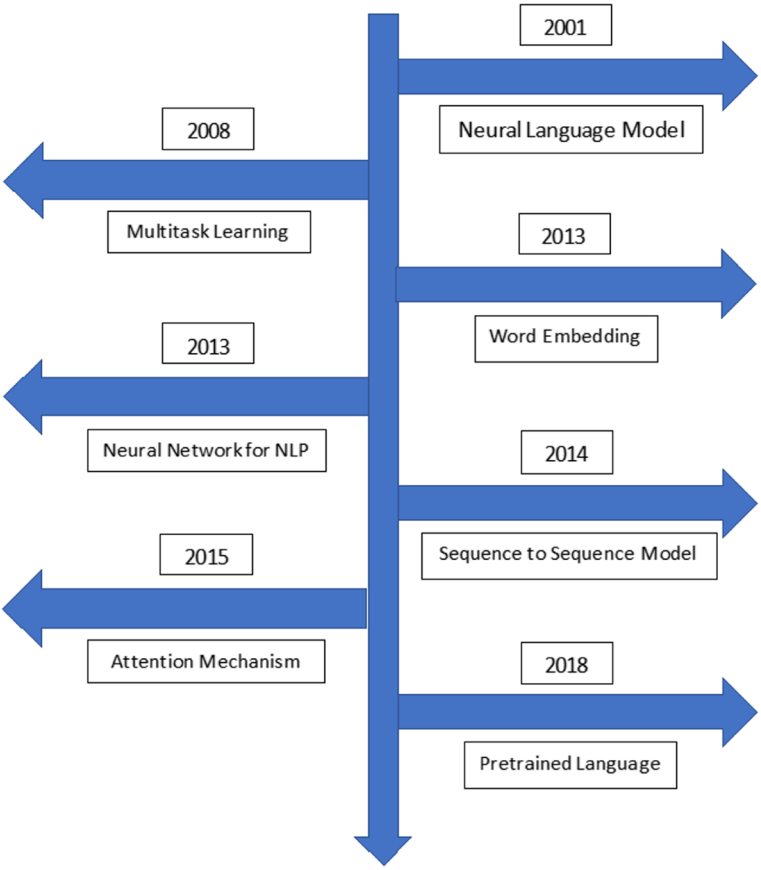

Fig. 3 A walkthrough of recent developments in NLP

Neural Networks those are recurrent in nature due to performing the same function for every data, also known as Recurrent Neural Networks (RNNs), have also been used in NLP, and found ideal for sequential data such as text, time series, financial data, speech, audio, video among others, see article by Thomas (2019) [137]. One of the modified versions of RNNs is Long Short-Term Memory (LSTM) which is also very useful in the cases where only the desired important information needs to be retained for a much longer time discarding the irrelevant information, see [52, 58]. Further development in the LSTM has also led to a slightly simpler variant, called the gated recurrent unit (GRU), which has shown better results than standard LSTMs in many tasks [22, 26]. *Attention mechanisms* [7] which suggest a network to learn what to pay attention to in accordance with the current hidden state and annotation together with the use of transformers have also made a significant development in NLP, see [141]. It is to be noticed that Transformers have a potential of learning longer-term dependency but are limited by a fixed-length context in the setting of language modeling. In this direction recently Dai et al. [30] proposed a novel neural architecture Transformer-XL (XL as extra-long) which enables learning dependencies beyond a fixed length of words. Further the work of Rae et al. [104] on the Compressive Transformer, an attentive sequence model which compresses memories for long-range sequence learning, may be helpful for the readers. One may also refer to the recent work by Otter et al. [98] on uses of Deep Learning for NLP, and

relevant references cited therein. The use of BERT (Bidirectional Encoder Representations from Transformers) [33] model and successive models have also played an important role for NLP.

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, Semantic Role Labeling have a huge contribution made to NLP, and are good topics for research. Sentiment analysis (Nasukawaetal.,2003) [156] works by extracting sentiments about a given topic, and it consists of a topic specific feature term extraction, sentiment extraction, and association by relationship analysis. It utilizes two linguistic resources for the analysis: the sentiment lexicon and the sentiment pattern database. It analyzes the documents for positive and negative words and tries to give ratings on scale −5 to +5. The mainstream of currently used tagsets is obtained from English. The most widely used tagsets as standard guidelines are designed for Indo-European languages but it is less researched on Asian languages or middle- eastern languages. Various authors have done research on making parts of speech taggers for various languages such as Arabic (Zeroual et al., 2017) [160], Sanskrit (Tapswi & Jain, 2012) [136], Hindi (Ranjan & Basu, 2003) [105] to efficiently tag and classify words as nouns, adjectives, verbs etc. Authors in [136] have used treebank technique for creating rule-based POS Tagger for Sanskrit Language. Sanskrit sentences are parsed to assign the appropriate tag to each word using suffix stripping algorithm, wherein the longest suffix is searched from the suffix table and tags are assigned. Diab et al. (2004) [34] used supervised machine learning approach and adopted Support Vector Machines (SVMs) which were trained on the Arabic Treebank to automatically tokenize parts of speech tag and annotate base phrases in Arabic text.

Chunking is a process of separating phrases from unstructured text. Since simple tokens may not represent the actual meaning of the text, it is advisable to use phrases such as “North Africa” as a single word instead of ‘North’ and ‘Africa’ separate words. Chunking known as “Shadow Parsing” labels parts of sentences with syntactic correlated keywords like Noun Phrase (NP) and Verb Phrase (VP). Chunking is often evaluated using the CoNLL 2000 shared task. Various researchers (Sha and Pereira, 2003; McDonald et al., 2005; Sun et al., 2008) [83, 122, 130] used CoNLL test data for chunking and used features composed of words, POS tags, and tags.

There are particular words in the document that refer to specific entities or real-world objects like location, people, organizations etc. To find the words which have a unique context and are more informative, noun phrases are considered in the text documents. Named entity recognition (NER) is a technique to recognize and separate the named entities and group them under predefined classes. But in the era of the Internet, where people use slang not the traditional or standard English which cannot be processed by standard natural language processing tools. Ritter (2011) [111] proposed the classification of named entities in tweets because standard NLP tools did not perform well on tweets. They re-built NLP pipeline starting from PoS tagging, then chunking for NER. It improved the performance in comparison to standard NLP tools.

Emotion detection investigates and identifies the types of emotion from speech, facial expressions, gestures, and text. Sharma (2016) [124] analyzed the conversations in Hinglish means mix of English and Hindi languages and identified the usage patterns of PoS. Their work was based on identification of language and POS tagging of mixed script. They tried to detect emotions in mixed script by relating machine learning and human knowledge. They have categorized sentences into 6 groups based on emotions and used TLBO technique to help

the users in prioritizing their messages based on the emotions attached with the message. Seal et al. (2020) [120] proposed an efficient emotion detection method by searching emotional words from a pre-defined emotional keyword database and analyzing the emotion words, phrasal verbs, and negation words. Their proposed approach exhibited better performance than recent approaches.

Semantic Role Labeling (SRL) works by giving a semantic role to a sentence. For example, in the PropBank (Palmer et al., 2005) [100] formalism, one assigns roles to words that are arguments of a verb in the sentence. The precise arguments depend on the verb frame and if multiple verbs exist in a sentence, it might have multiple tags. State-of-the-art SRL systems comprise several stages: creating a parse tree, identifying which parse tree nodes represent the arguments of a given verb, and finally classifying these nodes to compute the corresponding SRL tags.

Event discovery in social media feeds (Benson et al.,2011) [13], using a graphical model to analyze any social media feeds to determine whether it contains the 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 a broader array of features on factors.

We first give insights on some of the mentioned tools and relevant work done before moving to the broad applications of NLP.

3.2 Applications of NLP

Natural Language Processing can be applied into various areas like Machine Translation, Email Spam detection, Information Extraction, Summarization, Question Answering etc. Next, we discuss some of the areas with the relevant work done in those directions.

a) Machine Translation

As most of the world is online, the task of making data accessible and available to all is a challenge. Major challenge in making data accessible is the language barrier. There are a multitude of languages with different sentence structure and grammar. Machine Translation is generally translating phrases 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. The statistical machine learning gathers as many data as they can find that seems to be parallel between two languages and they crunch their data to find the likelihood that something in Language A corresponds to something in Language B. As for Google, in September 2016, announced a new machine translation system based on artificial neural networks and Deep learning. In recent years, various methods have been proposed to automatically evaluate machine translation quality by comparing hypothesis translations with reference translations. Examples of such methods are word error rate, position-independent word error rate (Tillmann et al., 1997) [138], generation string accuracy (Bangalore et al., 2000) [8], multi-reference word error rate (Nießen et al., 2000) [95], BLEU score (Papineni et al., 2002) [101], NIST score (Doddington, 2002) [35] All these criteria try to approximate human assessment and often achieve an astonishing degree of correlation to human subjective evaluation of fluency and adequacy (Papineni et al., 2001; Doddington, 2002) [35, 101].

b) Text Categorization

Categorization systems input a large flow of data like official documents, military casualty reports, market data, newswires etc. and assign them to predefined categories or indices. For example, The Carnegie Group's Construe system (Hayes, 1991) [54], inputs Reuters articles and saves much time by doing the work that is to be done by staff or human indexers. Some companies have been using categorization systems to categorize trouble tickets or complaint requests and routing to the appropriate desks. Another application of text categorization is email spam filters. Spam filters are becoming important as the first line of defence against the unwanted emails. A false negative and false positive issue of spam filters is at the heart of NLP technology, it has brought down the challenge of extracting meaning from strings of text. 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. There are several types of spam filters available. *Content filters*: Review the content within the message to determine whether it is spam or not. *Header filters*: Review the email header looking for fake information. *General Blacklist filters*: Stop all emails from blacklisted recipients. *Rules Based Filters*: It uses user-defined criteria. Such as stopping mails from a specific person or stopping mail including a specific word. *Permission Filters*: Require anyone sending a message to be pre-approved by the recipient. *Challenge Response Filters*: Requires anyone sending a message to enter a code to gain permission to send email.

c) 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 like Rule Learning (Cohen 1996) [27], Naïve Bayes (Sahami et al., 1998; Androutsopoulos et al., 2000; Rennie.,2000) [5, 109, 115],Memory based Learning (Sakkiset al.,2000b) [117], Support vector machines (Druker et al., 1999) [36], Decision Trees (Carreras and Marquez, 2001) [19], Maximum Entropy Model (Berger et al. 1996) [14], Hash Forest and a rule encoding method (T. Xia, 2020) [153], sometimes combining different learners (Sakkis et al., 2001) [116]. Using these approaches is better as classifier is learned from training data rather than making by hand. The naïve bayes is preferred because of its performance despite its simplicity (Lewis, 1998) [67] In Text Categorization two types of models have been used (McCallum and Nigam, 1998) [77]. Both modules assume that a fixed vocabulary is present. But in first model a document is generated by first choosing a subset of vocabulary and then using the selected words any number of times, at least once irrespective of order. This is called Multi-variate Bernoulli model. It takes the information of which words are used in a document irrespective of number of words and order. In second model, a document is generated by choosing a set of word occurrences and arranging them in any order. This model is called multi-nomial model, in addition to the Multi-variate Bernoulli model, it also captures information on how many times a word is used in a document. Most text categorization approaches to anti-spam Email filtering have used multi variate Bernoulli model (Androutsopoulos et al., 2000) [5] [15].

d) 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 summarizing the information relevant to a user's needs. In the case of a domain specific search engine, the automatic identification of important information can increase accuracy and efficiency of a directed search. There is use of hidden Markov models (HMMs) to extract the relevant fields of research papers. These extracted text segments are used to allow searched over specific fields and to provide effective presentation of search results and to match references to papers. For example, noticing the pop-up ads on any websites showing the recent items you might have looked on an online store with discounts. In Information Retrieval two types of models have been used (McCallum and Nigam, 1998) [77]. Both modules assume that a fixed vocabulary is present. But in first model a document is generated by first choosing a subset of vocabulary and then using the selected words any number of times, at least once without any order. This is called Multi-variate Bernoulli model. It takes the information of which words are used in a document irrespective of number of words and order. In second model, a document is generated by choosing a set of word occurrences and arranging them in any order. This model is called multi-nominal model, in addition to the Multi-variate Bernoulli model, it also captures information on how many times a word is used in a document.

Discovery of knowledge is becoming important areas of research over the recent years. Knowledge discovery research use a variety of techniques to extract useful information from source documents like *Parts of Speech (POS) tagging*, *Chunking or Shadow Parsing*, *Stop-words* (Keywords that are used and must be removed before processing documents), *Stemming* (Mapping words to some base for, it has two methods, dictionary-based stemming and Porter style stemming (Porter, 1980) [103]. Former one has higher accuracy but higher cost of implementation while latter has lower implementation cost and is usually insufficient for IR). *Compound or Statistical Phrases* (Compounds and statistical phrases index multi token units instead of single tokens.) *Word Sense Disambiguation* (Word sense disambiguation is the task of understanding the correct sense of a word in context. When used for information retrieval, terms are replaced by their senses in the document vector.)

The extracted information can be applied for a variety of purposes, for example to prepare a summary, to build databases, identify keywords, classifying text items according to some pre-defined categories etc. For example, CONSTRUE, it was developed for Reuters, that is used in classifying news stories (Hayes, 1992) [54]. It has been suggested that many IE systems can successfully extract terms from documents, acquiring relations between the terms is still a difficulty. PROMETHEE is a system that extracts lexico-syntactic patterns relative to a specific conceptual relation (Morin,1999) [89]. IE systems should work at many levels, from word recognition to discourse analysis at the level of the complete document. An application of the Blank Slate Language Processor (BSLP) (Bondale et al., 1999) [16] approach for the analysis of a real-life natural language corpus that consists of responses to open-ended questionnaires in the field of advertising.

There is a system called MITA (Metlife's Intelligent Text Analyzer) (Glasgow et al. (1998) [48]) that extracts information from life insurance applications. Ahonen et al. (1998) [1] suggested a mainstream framework for text mining that uses pragmatic and discourse level analyses of text.

e) Summarization

Overload of information is the real thing in this digital age, and already our reach and access to knowledge and information exceeds our capacity to understand it. This trend is not slowing

down, so an ability to summarize the data while keeping the meaning intact is highly required. This is important not just allowing us the ability to recognize the understand the important information for a large set of data, it is used to understand the deeper emotional meanings; For example, a company determines the general sentiment on social media and uses it on their latest product offering. This application is useful as a valuable marketing asset.

The types of text summarization depends on the basis of the number of documents and the two important categories are single document summarization and multi document summarization (Zajic et al. 2008 [159]; Fattah and Ren 2009 [43]).Summaries can also be of two types: generic or query-focused (Gong and Liu 2001 [50]; Dunlavy et al. 2007 [37]; Wan 2008 [144]; Ouyang et al. 2011 [99]).Summarization task can be either supervised or unsupervised (Mani and Maybury 1999 [74]; Fattah and Ren 2009 [43]; Riedhammer et al. 2010 [110]). Training data is required in a supervised system for selecting relevant material from the documents. Large amount of annotated data is needed for learning techniques. Few techniques are as follows–

- *Bayesian Sentence based Topic Model (BSTM)* uses both term-sentences and term document associations for summarizing multiple documents. (Wang et al. 2009 [146])
- *Factorization with Given Bases (FGB)* is a language model where sentence bases are the given bases and it utilizes document-term and sentence term matrices. This approach groups and summarizes the documents simultaneously. (Wang et al. 2011) [147])
- *Topic Aspect-Oriented Summarization (TAOS)* is based on topic factors. These topic factors are various features that describe topics such as capital words are used to represent entity. Various topics can have various aspects and various preferences of features are used to represent various aspects. (Fang et al. 2015 [42])

f) Dialogue System

Dialogue systems are very prominent in real world applications ranging from providing support to performing a particular action. In case of support dialogue systems, context awareness is required whereas in case to perform an action, it doesn't require much context awareness. Earlier dialogue systems were focused on small applications such as home theater systems. These dialogue systems utilize phonemic and lexical levels of language. Habitable dialogue systems offer potential for fully automated dialog systems by utilizing all levels of a language. (Liddy, 2001) [68].This leads to producing systems that can enable robots to interact with humans in natural languages such as Google's assistant, Windows Cortana, Apple's Siri and Amazon's Alexa etc.

g) Medicine

NLP is applied in the field as well. The Linguistic String Project-Medical Language Processor is one the large scale projects of NLP in the field of medicine [21, 53, 57, 71, 114]. The LSP-MLP helps enabling physicians to extract and summarize information of any signs or symptoms, drug dosage and response data with the aim of identifying possible side effects of any medicine while highlighting or flagging data items [114]. The National Library of Medicine is developing The Specialist System [78–80, 82, 84]. It is expected to function as an Information Extraction tool for Biomedical Knowledge Bases, particularly Medline abstracts.

The lexicon was created using MeSH (Medical Subject Headings), Dorland's Illustrated Medical Dictionary and general English Dictionaries. The Centre d'Informatique Hospitaliere of the Hopital Cantonal de Geneve is working on an electronic archiving environment with NLP features [81, 119]. In the first phase, patient records were archived. At later stage the LSP-MLP has been adapted for French [10, 72, 94, 113], and finally, a proper NLP system called RECIT [9, 11, 17, 106] has been developed using a method called Proximity Processing [88]. It's task was to implement a robust and multilingual system able to analyze/comprehend medical sentences, and to preserve a knowledge of free text into a language independent knowledge representation [107, 108]. The Columbia university of New York has developed an NLP system called MEDLEE (MEDical Language Extraction and Encoding System) that identifies clinical information in narrative reports and transforms the textual information into structured representation [45].

3.3 NLP in talk

We next discuss some of the recent NLP projects implemented by various companies:

a) ACE Powered GDPR Robot Launched by RAVN Systems [134]

RAVN Systems, a 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. The Robot uses AI techniques to automatically analyze documents and other types of data in any business system which is subject to GDPR rules. It allows users to search, retrieve, flag, classify, and report on data, mediated to be super sensitive under GDPR quickly and easily. Users also can identify personal data from documents, view feeds on the latest personal data that requires attention and provide reports on the data suggested to be deleted or secured. RAVN's GDPR Robot is also able to hasten requests for information (Data Subject Access Requests - "DSAR") in a simple and efficient way, removing the need for a physical approach to these requests which tends to be very labor thorough. Peter Wallqvist, CSO at RAVN Systems commented, "GDPR compliance is of universal paramountcy as it will be exploited by any organization that controls and processes data concerning EU citizens.

Link: http://markets.financialcontent.com/stocks/news/read/33888795/RAVN_Systems_Launch_the_ACE_Powered_GDPR_Robot

b) Eno A Natural Language Chatbot Launched by Capital One [56]

Capital One announces a chatbot for customers called Eno. Eno is a natural language chatbot that people socialize through texting. CapitalOne claims that Eno is First natural language SMS chatbot from a U.S. bank that allows customers 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. This provides a different platform than other brands that launch chatbots like Facebook Messenger and Skype. They believed that Facebook has too much access to private information of a person, which could get them into trouble with privacy laws U.S. financial institutions work under. Like Facebook Page admin can access full transcripts of the bot's conversations. If that

would be the case then the admins could easily view the personal banking information of customers with is not correct.

Link: https://www.macobserver.com/analysis/capital-one-natural-language-chatbot-eno/

c) Future of BI in Natural Language Processing [140]

Several companies in BI spaces are trying to get with the trend and trying hard to ensure that data becomes more friendly and easily accessible. But still there is a long way for this.BI will also make it easier to access as GUI is not needed. Because nowadays the queries are made by text or voice command on smartphones.one of the most common examples is Google might tell you today what tomorrow's weather will be. But soon enough, we will be able to ask our personal data chatbot about customer sentiment today, and how 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. Initially, the data chatbot will probably ask the question 'how have revenues changed over the last three-quarters?' and then return pages of data for you to analyze. But once it learns the semantic relations and inferences of the question, it will be able to automatically perform the filtering and formulation necessary to provide an intelligible answer, rather than simply showing you data.

Link: http://www.smartdatacollective.com/eran-levy/489410/here-s-why-natural-language-processing-future-bi

d) Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research [97]

Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research describes a theory derivation process that is used to develop a conceptual framework for medication therapy management (MTM) research. The MTM service model and chronic care model are selected as parent theories. Review article abstracts target medication therapy management in chronic disease care that were retrieved from Ovid Medline (2000–2016). Unique concepts in each abstract are extracted using Meta Map and their pair-wise co-occurrence are determined. Then the information is used to construct a network graph of concept co-occurrence that is further analyzed to identify content for the new conceptual model. 142 abstracts are analyzed. Medication adherence is the most studied drug therapy problem and co-occurred with concepts related to patient-centered interventions targeting self-management. The enhanced model consists of 65 concepts clustered into 14 constructs. The framework requires additional refinement and evaluation to determine its relevance and applicability across a broad audience including underserved settings.

Link: https://www.ncbi.nlm.nih.gov/pubmed/28269895?dopt=Abstract

e) Meet the Pilot, world's first language translating earbuds [96]

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. Simultaneously, the user will hear the translated version of the speech on the second earpiece. Moreover, it is not necessary that conversation would be taking place between two people; only the users can join in and discuss as a group. As if now the user may experience a few second lag interpolated the speech and translation, which Waverly Labs pursue to reduce. The Pilot earpiece will be available from September but can be pre-ordered now for $249. The earpieces can also be used for streaming music, answering voice calls, and getting audio notifications.

Link: https://www.indiegogo.com/projects/meet-the-pilot-smart-earpiece-language-translator-headphones-travel#/

4 Datasets in NLP and state-of-the-art models

The objective of this section is to present the various datasets used in NLP and some state-of-the-art models in NLP.

4.1 Datasets in NLP

Corpus is a collection of linguistic data, either compiled from written texts or transcribed from recorded speech. Corpora are intended primarily for testing linguistic hypotheses - e.g., to determine how a certain sound, word, or syntactic construction is used across a culture or language. There are various types of corpus: In an annotated corpus, the implicit information in the plain text has been made explicit by specific annotations. Un-annotated corpus contains raw state of plain text. Different languages can be compared using a reference corpus. Monitor corpora are non-finite collections of texts which are mostly used in lexicography. Multilingual corpus refers to a type of corpus that contains small collections of monolingual corpora based on the same sampling procedure and categories for different languages. Parallel corpus contains texts in one language and their translations into other languages which are aligned sentence phrase by phrase. Reference corpus contains text of spoken (formal and informal) and written (formal and informal) language which represents various social and situational contexts. Speech corpus contains recorded speech and transcriptions of recording and the time each word occurred in the recorded speech. There are various datasets available for natural language processing; some of these are listed below for different use cases:

1. Sentiment Analysis: Sentiment analysis is a rapidly expanding field of natural language processing (NLP) used in a variety of fields such as politics, business etc. Majorly used datasets for sentiment analysis are:

a) Stanford Sentiment Treebank (SST): Socher et al. introduced SST containing sentiment labels for 215,154 phrases in parse trees for 11,855 sentences from movie reviews posing novel sentiment compositional difficulties [127].
b) Sentiment140: It contains 1.6 million tweets annotated with negative, neutral and positive labels.

c) Paper Reviews: It provides reviews of computing and informatics conferences written in English and Spanish languages. It has 405 reviews which are evaluated on a 5-point scale ranging from very negative to very positive.
d) IMDB: For natural language processing, text analytics, and sentiment analysis, this dataset offers thousands of movie reviews split into training and test datasets. This dataset was introduced in by Mass et al. in 2011 [73].
e) G.Rama Rohit Reddy of the Language Technologies Research Centre, KCIS, IIIT Hyderabad, generated the corpus "Sentiraama." The corpus is divided into four datasets, each of which is annotated with a two-value scale that distinguishes between positive and negative sentiment at the document level. The corpus contains data from a variety of fields, including book reviews, product reviews, movie reviews, and song lyrics. The annotators meticulously followed the annotation technique for each of them. The folder "Song Lyrics" in the corpus contains 339 Telugu song lyrics written in Telugu script [121].

2. Language Modelling: Language models analyse text data to calculate word probability. They use an algorithm to interpret the data, which establishes rules for context in natural language. The model then uses these rules to accurately predict or construct new sentences. The model basically learns the basic characteristics and features of language and then applies them to new phrases. Majorly used datasets for Language modeling are as follows:

a) Salesforce's WikiText-103 dataset has 103 million tokens collected from 28,475 featured articles from Wikipedia.
b) WikiText-2 is a scaled-down version of WikiText-103. It contains 2 million tokens with a 33,278 jargon size.
c) Penn Treebank piece of the Wall Street Diary corpus includes 929,000 tokens for training, 73,000 tokens for validation, and 82,000 tokens for testing purposes. Its context is limited since it comprises sentences rather than paragraphs [76].
d) The Ministry of Electronics and Information Technology's Technology Development Programme for Indian Languages (TDIL) launched its own data distribution portal (www.tdil-dc.in) which has cataloged datasets [24].

3. Machine Translation: The task of converting the text of one natural language into another language while keeping the sense of the input text is known as machine translation. Majorly used datasets are as follows:

a) Tatoeba is a collection of multilingual sentence pairings. A tab-delimited pair of an English text sequence and the translated French text sequence appears on each line of the dataset. Each text sequence might be as simple as a single sentence or as complex as a paragraph of many sentences.
b) The Europarl parallel corpus is derived from the European Parliament's proceedings. It is available in 21 European languages [40].
c) WMT14 provides machine translation pairs for English-German and English-French. Separately, these datasets comprise 4.5 million and 35 million sentence sets. Byte-Pair Encoding with 32 K tasks is used to encode the phrases.

d) There are around 160,000 sentence pairings in the IWSLT 14. The dataset includes descriptions in English-German (En-De) and German-English (De-En) languages. There are around 200 K training sentence sets in the IWSLT 13 dataset.
e) The IIT Bombay English-Hindi corpus comprises parallel corpora for English-Hindi as well as monolingual Hindi corpora gathered from several existing sources and corpora generated over time at IIT Bombay's Centre for Indian Language Technology.

4. Question Answering System: Question answering systems provide real-time responses which are widely used in customer care services. The datasets used for dialogue system/ question answering system are as follows:

a) Stanford Question Answering Dataset (SQuAD): it is a reading comprehension dataset made up of questions posed by crowd workers on a collection of Wikipedia articles.
b) Natural Questions: It is a large-scale corpus presented by Google used for training and assessing open-domain question answering systems. It includes 300,000 naturally occurring queries as well as human-annotated responses from Wikipedia pages for use in QA system training.
c) Question Answering in Context (QuAC): This dataset is used to describe, comprehend, and participate in information seeking conversation. In this dataset, instances are made up of an interactive discussion between two crowd workers: a student who asks a series of open-ended questions about an unknown Wikipedia text, and a teacher who responds by offering brief extracts from the text.

The neural learning models are overtaking traditional models for NLP [64, 127]. In [64], authors used CNN (Convolutional Neural Network) model for sentiment analysis of movie reviews and achieved 81.5% accuracy. The results illustrate that using CNN was an appropriate replacement for state-of-the-art methods. Authors [127] have combined SST and Recursive Neural Tensor Network for sentiment analysis of the single sentence. This model amplifies the accuracy by 5.4% for sentence classification compared to traditional NLP models. Authors [135] proposed a combined Recurrent Neural Network and Transformer model for sentiment analysis. This hybrid model was tested on three different datasets: Twitter US Airline Sentiment, IMDB, and Sentiment 140: and achieved F1 scores of 91%, 93%, and 90%, respectively. This model's performance outshined the state-of-art methods.

Santoro et al. [118] introduced a rational recurrent neural network with the capacity to learn on classifying the information and perform complex reasoning based on the interactions between compartmentalized information. They used the relational memory core to handle such interactions. Finally, the model was tested for language modeling on three different datasets (GigaWord, Project Gutenberg, and WikiText-103). Further, they mapped the performance of their model to traditional approaches for dealing with relational reasoning on compartmentalized information. The results achieved with RMC show improved performance.

Merity et al. [86] extended conventional word-level language models based on Quasi-Recurrent Neural Network and LSTM to handle the granularity at character and word level. They tuned the parameters for character-level modeling using Penn Treebank dataset and word-level modeling using WikiText-103. In both cases, their model outshined the state-of-art methods.

Luong et al. [70] used neural machine translation on the WMT14 dataset and performed translation of English text to French text. The model demonstrated a significant improvement

of up to 2.8 bi-lingual evaluation understudy (BLEU) scores compared to various neural machine translation systems. It outperformed the commonly used MT system on a WMT 14 dataset.

Fan et al. [41] introduced a gradient-based neural architecture search algorithm that automatically finds architecture with better performance than a transformer, conventional NMT models. They tested their model on WMT14 (English-German Translation), IWSLT14 (German-English translation), and WMT18 (Finnish-to-English translation) and achieved 30.1, 36.1, and 26.4 BLEU points, which shows better performance than Transformer baselines.

Wiese et al. [150] introduced a deep learning approach based on domain adaptation techniques for handling biomedical question answering tasks. Their model revealed the state-of-the-art performance on biomedical question answers, and the model outperformed the state-of-the-art methods in domains.

Seunghak et al. [158] designed a Memory-Augmented-Machine-Comprehension-Network (MAMCN) to handle dependencies faced in reading comprehension. The model achieved state-of-the-art performance on document-level using TriviaQA and QUASAR-T datasets, and paragraph-level using SQuAD datasets.

Xie et al. [154] proposed a neural architecture where candidate answers and their representation learning are constituent centric, guided by a parse tree. Under this architecture, the search space of candidate answers is reduced while preserving the hierarchical, syntactic, and compositional structure among constituents. Using SQuAD, the model delivers state-of-the-art performance.

4.2 State-of-the-art models in NLP

Rationalist approach or symbolic approach assumes that a crucial part of the knowledge in the human mind is not derived by the senses but is firm in advance, probably by genetic inheritance. Noam Chomsky was the strongest advocate of this approach. It was believed that machines can be made to function like the human brain by giving some fundamental knowledge and reasoning mechanism linguistics knowledge is directly encoded in rule or other forms of representation. This helps the automatic process of natural languages [92]. Statistical and machine learning entail evolution of algorithms that allow a program to infer patterns. An iterative process is used to characterize a given algorithm's underlying algorithm that is optimized by a numerical measure that characterizes numerical parameters and learning phase. Machine-learning models can be predominantly categorized as either generative or discriminative. Generative methods can generate synthetic data because of which they create rich models of probability distributions. Discriminative methods are more functional and have right estimating posterior probabilities and are based on observations. Srihari [129] explains the different generative models as one with a resemblance that is used to spot an unknown speaker's language and would bid the deep knowledge of numerous languages to perform the match. Discriminative methods rely on a less knowledge-intensive approach and using distinction between languages. Whereas generative models can become troublesome when many features are used and discriminative models allow use of more features [38]. Few of the examples of discriminative methods are Logistic regression and conditional random fields (CRFs), generative methods are Naive Bayes classifiers and hidden Markov models (HMMs).

(a) Naive Bayes Classifiers

Naive Bayes is a probabilistic algorithm which is based on probability theory and Bayes' Theorem to predict the tag of a text such as news or customer review. It helps to calculate the probability of each tag for the given text and return the tag with the highest probability. Bayes' Theorem is used to predict the probability of a feature based on prior knowledge of conditions that might be related to that feature. The choice of area in NLP using Naïve Bayes Classifiers could be in usual tasks such as segmentation and translation but it is also explored in unusual areas like segmentation for infant learning and identifying documents for opinions and facts. Anggraeni et al. (2019) [61] used ML and AI to create a question-and-answer system for retrieving information about hearing loss. They developed I-Chat Bot which understands the user input and provides an appropriate response and produces a model which can be used in the search for information about required hearing impairments. The problem with naïve bayes is that we may end up with zero probabilities when we meet words in the test data for a certain class that are not present in the training data.

(b) 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. We can describe the outputs, but the system's internals are hidden. Few of the problems could be solved by Inference A certain sequence of output symbols, compute the probabilities of one or more candidate states with sequences. *Patterns matching* the state-switch sequence 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, where the output sequence is matched to the sequence of individual phonemes. HMM is not restricted to this application; it has several others such as bioinformatics problems, for example, multiple sequence alignment [128]. Sonnhammer mentioned that Pfam holds multiple alignments and hidden Markov model-based profiles (HMM-profiles) of entire protein domains. The cue of domain boundaries, family members and alignment are done semi-automatically found on expert knowledge, sequence similarity, other protein family databases and the capability of HMM-profiles to correctly identify and align the members. HMM may be used for a variety of NLP applications, including word prediction, sentence production, quality assurance, and intrusion detection systems [133].

(c) Neural Network

Earlier machine learning techniques such as Naïve Bayes, HMM etc. were majorly used for NLP but by the end of 2010, neural networks transformed and enhanced NLP tasks by learning multilevel features. Major use of neural networks in NLP is observed for word embedding where words are represented in the form of vectors. These vectors can be used to recognize similar words by observing their closeness in this vector space, other uses of neural networks are observed in information retrieval, text summarization, text classification, machine translation, sentiment analysis and speech recognition. Initially focus was on feedforward [49] and CNN (convolutional neural network) architecture [69] but later researchers adopted recurrent neural networks to capture the context of a word with respect to surrounding words of a sentence. LSTM (Long Short-Term Memory), a variant of RNN, is used in various tasks such

as word prediction, and sentence topic prediction. [47] In order to observe the word arrangement in forward and backward direction, bi-directional LSTM is explored by researchers [59]. In case of machine translation, encoder-decoder architecture is used where dimensionality of input and output vector is not known. Neural networks can be used to anticipate a state that has not yet been seen, such as future states for which predictors exist whereas HMM predicts hidden states.

(d) BERT

Bi-directional Encoder Representations from Transformers (BERT) is a pre-trained model with unlabeled text available on BookCorpus and English Wikipedia. This can be fine-tuned to capture context for various NLP tasks such as question answering, sentiment analysis, text classification, sentence embedding, interpreting ambiguity in the text etc. [25, 33, 90, 148]. Earlier language-based models examine the text in either of one direction which is used for sentence generation by predicting the next word whereas the BERT model examines the text in both directions simultaneously for better language understanding. BERT provides contextual embedding for each word present in the text unlike context-free models (word2vec and GloVe). For example, in the sentences "he is going to the riverbank for a walk" and "he is going to the bank to withdraw some money", word2vec will have one vector representation for "bank" in both the sentences whereas BERT will have different vector representation for "bank". Muller et al. [90] used the BERT model to analyze the tweets on covid-19 content. The use of the BERT model in the legal domain was explored by Chalkidis et al. [20].

Since BERT considers up to 512 tokens, this is the reason if there is a long text sequence that must be divided into multiple short text sequences of 512 tokens. This is the limitation of BERT as it lacks in handling large text sequences.

5 Evaluation metrics and challenges

The objective of this section is to discuss evaluation metrics used to evaluate the model's performance and involved challenges.

5.1 Evaluation metrics

Since the number of labels in most classification problems is fixed, it is easy to determine the score for each class and, as a result, the loss from the ground truth. In image generation problems, the output resolution and ground truth are both fixed. As a result, we can calculate the loss at the pixel level using ground truth. But in NLP, though output format is predetermined in the case of NLP, dimensions cannot be specified. It is because a single statement can be expressed in multiple ways without changing the intent and meaning of that statement. Evaluation metrics are important to evaluate the model's performance if we were trying to solve two problems with one model.

a) BLEU (BiLingual Evaluation Understudy) Score: Each word in the output sentence is scored 1 if it appears in either of the reference sentences and a 0 if it does not. Further the number of words that appeared in one of the reference translations is divided by the total number of words in the output sentence to normalize the count so that it is always between

0 and 1. For example, if ground truth is "He is playing chess in the backyard" and output sentences are S1: "He is playing tennis in the backyard", S2: "He is playing badminton in the backyard", S3: "He is playing movie in the backyard" and S4: "backyard backyard backyard backyard backyard backyard backyard". The score of S1, S2 and S3 would be 6/7,6/7 and 6/7. All sentences are getting the same score though information in S1 and S3 is not same. This is because BELU considers words in a sentence contribute equally to the meaning of a sentence which is not the case in real-world scenario. Using combination of uni-gram, bi-gram and n-grams, we can to capture the order of a sentence. We may also set a limit on how many times each word is counted based on how many times it appears in each reference phrase, which helps us prevent excessive repetition.

b) GLUE (General Language Understanding Evaluation) score: Previously, NLP models were almost usually built to perform effectively on a unique job. Various models such as LSTM, Bi-LSTM were trained solely for this task, and very rarely generalized to other tasks. The model which is used for named entity recognition can perform for textual entailment. GLUE is a set of datasets for training, assessing, and comparing NLP models. It includes nine diverse task datasets designed to test a model's language understanding. To acquire a comprehensive assessment of a model's performance, GLUE tests the model on a variety of tasks rather than a single one. Single-sentence tasks, similarity and paraphrase tasks, and inference tasks are among them. For example, in sentiment analysis of customer reviews, we might be interested in analyzing ambiguous reviews and determining which product the client is referring to in his reviews. Thus, the model obtains a good "knowledge" of language in general after some generalized pre-training. When the time comes to test out a model to meet a given task, this universal "knowledge" gives us an advantage. With GLUE, researchers can evaluate their model and score it on all nine tasks. The final performance score model is the average of those nine scores. It makes little difference how the model looks or works if it can analyze inputs and predict outcomes for all the activities.

Considering these metrics in mind, it helps to evaluate the performance of an NLP model for a particular task or a variety of tasks.

5.2 Challenges

The applications of NLP have been growing day by day, and with these new challenges are also occurring despite a lot of work done in the recent past. Some of the common challenges are: Contextual words and phrases in the language where same words and phrases can have different meanings in a sentence which are easy for the humans to understand but makes a challenging task. Such type of challenges can also be faced with dealing Synonyms in the language because humans use many different words to express the same idea, also in the language different levels of complexity such as large, huge, and big may be used by the different persons which become a challenging task to process the language and design algorithms to adopt all these issues. Further in language, Homonyms, the words used to be pronounced the same but have different definitions are also problematic for question answering and speech-to-text applications because they aren't written in text form. Sentences using sarcasm and irony sometimes may be understood in the opposite way by the humans, and so designing models to deal with such sentences is a really challenging task in NLP. Furthermore, the sentences in the language having any type of ambiguity in the sense of interpreting in more

than one way is also an area to work upon where more accuracy can be achieved. Language containing informal phrases, expressions, idioms, and culture-specific lingo make difficult to design models intended for the broad use, however having a lot of data on which training and updating on regular basis may improve the models, but it is a really challenging task to deal with the words having different meaning in different geographic areas. In fact, such types of issues also occur in dealing with different domains such as the meaning of words or sentences may be different in the education industry but have different meaning in health, law, defense etc. So, the models for NLP may be working good for an individual domain, geographic area but for a broad use such challenges need to be tackled. Further together with the above-mentioned challenges misspelled or misused words can also create a problem, although autocorrect and grammar corrections applications have improved a lot due to the continuous developments in the direction but predicting the intention of the writer that to from a specific domain, geographic area by considering sarcasm, expressions, informal phrases etc. is really a big challenge. There is no doubt that for most common widely used languages models for NLP have been doing very well, and further improving day by day but still there is a need for models for all the persons rather than specific knowledge of a particular language and technology. One may further refer to the work of Sharifirad and Matwin (2019) [123] for classification of different online harassment categories and challenges, Baclic et.al. (2020) [6] and Wong et al. (2018) [151] for challenges and opportunities in public health, Kang et.al. (2020) [63] for detailed literature survey and technological challenges relevant to management research and NLP, and a recent review work by Alshemali and Kalita (2020) [3], and references cited there in.

In the recent past, models dealing with Visual Commonsense Reasoning [31] and NLP have also been getting attention of the several researchers and seems a promising and challenging area to work upon. These models try to extract the information from an image, video using a visual reasoning paradigm such as the humans can infer from a given image, video beyond what is visually obvious, such as objects' functions, people's intents, and mental states. In this direction, recently Wen and Peng (2020) [149] suggested a model to capture knowledge from different perspectives, and perceive common sense in advance, and the results based on the conducted experiments on visual commonsense reasoning dataset VCR seems very satisfactory and effective. The work of Peng and Chi (2019) [102], that proposes Domain Adaptation with Scene Graph approach to transfer knowledge from the source domain with the objective to improve cross-media retrieval in the target domain, and Yen et al. (2019) [155] is also very useful to further explore the use of NLP and in its relevant domains.

6 Conclusion

This paper is written with three objectives. The first objective gives insights of the various important terminologies of NLP and NLG, and can be useful for the readers interested to start their early career in NLP and work relevant to its applications. The second objective of this paper focuses on the history, applications, and recent developments in the field of NLP. The third objective is to discuss datasets, approaches and evaluation metrics used in NLP. The relevant work done in the existing literature with their findings and some of the important applications and projects in NLP are also discussed in the paper. The last two objectives may serve as a literature survey for the readers already working in the NLP and relevant fields, and further can provide motivation to explore the fields mentioned in this paper. It is to be noticed

that even though a great amount of work on natural language processing is available in literature surveys (one may refer to [15, 32, 63, 98, 133, 151] focusing on one domain such as usage of deep-learning techniques in NLP, techniques used for email spam filtering, medication safety, management research, intrusion detection, and Gujarati language etc.), still there is not much work on regional languages, which can be the focus of future research.

Acknowledgements Authors would like to express the gratitude to Research Mentors from CL Educate: Accendere Knowledge Management Services Pvt. Ltd. for their comments on earlier versions of the manuscript. Although any errors are our own and should not tarnish the reputations of these esteemed persons. We would also like to appreciate the Editor, Associate Editor, and anonymous referees for their constructive suggestions that led to many improvements on an earlier version of this manuscript.

Declarations

Conflict of interest The first draft of this paper was written under the supervision of Dr. Kiran Khatter and Dr. Sukhdev Singh, associated with CL- Educate: Accendere Knowledge Management Services Pvt. Ltd. and deputed at the Manav Rachna International University. The draft is also available on arxiv.org at https://arxiv.org/abs/1708.05148

References

1. Ahonen H, Heinonen O, Klemettinen M, Verkamo AI (1998) Applying data mining techniques for descriptive phrase extraction in digital document collections. In research and technology advances in digital libraries, 1998. ADL 98. Proceedings. IEEE international forum on (pp. 2-11). IEEE
2. Alshawi H (1992) The core language engine. MIT press
3. Alshemali B, Kalita J (2020) Improving the reliability of deep neural networks in NLP: A review. Knowl-Based Syst 191:105210
4. Andreev ND (1967) The intermediary language as the focal point of machine translation. In: Booth AD (ed) Machine translation. North Holland Publishing Company, Amsterdam, pp 3–27
5. Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos CD, Stamatopoulos P (2000) Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009
6. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J (2020) Artificial intelligence in public health: challenges and opportunities for public health made possible by advances in natural language processing. Can Commun Dis Rep 46(6):161
7. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In ICLR 2015
8. Bangalore S, Rambow O, Whittaker S (2000) Evaluation metrics for generation. In proceedings of the first international conference on natural language generation-volume 14 (pp. 1-8). Assoc Comput Linguist
9. Baud RH, Rassinoux AM, Scherrer JR (1991) Knowledge representation of discharge summaries. In AIME 91 (pp. 173–182). Springer, Berlin Heidelberg
10. Baud RH, Rassinoux AM, Scherrer JR (1992) Natural language processing and semantical representation of medical texts. Methods Inf Med 31(2):117–125
11. Baud RH, Alpay L, Lovis C (1994) Let's meet the users with natural language understanding. Knowledge and Decisions in Health Telematics: The Next Decade 12:103
12. Bengio Y, Ducharme R, Vincent P (2001) A neural probabilistic language model. Proceedings of NIPS
13. Benson E, Haghighi A, Barzilay R (2011) Event discovery in social media feeds. In proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies-volume 1 (pp. 389-398). Assoc Comput Linguist
14. Berger AL, Della Pietra SA, Della Pietra VJ (1996) A maximum entropy approach to natural language processing. Computational Linguistics 22(1):39–71
15. Blanzieri E, Bryl A (2008) A survey of learning-based techniques of email spam filtering. Artif Intell Rev 29(1):63–92

16. Bondale N, Maloor P, Vaidyanathan A, Sengupta S, Rao PV (1999) Extraction of information from open-ended questionnaires using natural language processing techniques. Computer Science and Informatics 29(2):15–22
17. Borst F, Sager N, Nhàn NT, Su Y, Lyman M, Tick LJ, ..., Scherrer JR (1989) Analyse automatique de comptes rendus d'hospitalisation. In Degoulet P, Stephan JC, Venot A, Yvon PJ, rédacteurs. Informatique et Santé, Informatique et Gestion des Unités de Soins, Comptes Rendus du Colloque AIM-IF, Paris (pp. 246–56). [5]
18. Briscoe EJ, Grover C, Boguraev B, Carroll J (1987) A formalism and environment for the development of a large grammar of English. IJCAI 87:703–708
19. Carreras X, Marquez L (2001) Boosting trees for anti-spam email filtering. arXiv preprint cs/0109015
20. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the muppets straight out of law school. arXiv preprint arXiv:2010.02559
21. Chi EC, Lyman MS, Sager N, Friedman C, Macleod C (1985) A database of computer-structured narrative: methods of computing complex relations. In proceedings of the annual symposium on computer application in medical care (p. 221). Am Med Inform Assoc
22. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y, (2014) On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259
23. Chomsky N (1965) Aspects of the theory of syntax. MIT Press, Cambridge, Massachusetts
24. Choudhary N (2021) LDC-IL: the Indian repository of resources for language technology. Lang Resources & Evaluation 55:855–867. https://doi.org/10.1007/s10579-020-09523-3
25. Chouikhi H, Chniter H, Jarray F (2021) Arabic sentiment analysis using BERT model. In international conference on computational collective intelligence (pp. 621-632). Springer, Cham
26. Chung J, Gulcehre C, Cho K, Bengio Y, (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
27. Cohen WW (1996) Learning rules that classify e-mail. In AAAI spring symposium on machine learning in information access (Vol. 18, p. 25)
28. Cohen PR, Morgan J, Ramsay AM (2002) Intention in communication, Am J Psychol 104(4)
29. Collobert R, Weston J (2008) A unified architecture for natural language processing. In proceedings of the 25th international conference on machine learning (pp. 160–167)
30. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R, (2019) Transformer-xl: attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860
31. Davis E, Marcus G (2015) Commonsense reasoning and commonsense knowledge in artificial intelligence. Commun ACM 58(9):92–103
32. Desai NP, Dabhi VK (2022) Resources and components for Gujarati NLP systems: a survey. Artif Intell Rev:1–19
33. Devlin J, Chang MW, Lee K, Toutanova K, (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
34. Diab M, Hacioglu K, Jurafsky D (2004) Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short papers (pp. 149–152). Assoc Computat Linguist
35. Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In proceedings of the second international conference on human language technology research (pp. 138-145). Morgan Kaufmann publishers Inc
36. Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 10(5):1048–1054
37. Dunlavy DM, O'Leary DP, Conroy JM, Schlesinger JD (2007) QCS: A system for querying, clustering and summarizing documents. Inf Process Manag 43(6):1588–1605
38. Elkan C (2008) Log-Linear Models and Conditional Random Fields. http://cseweb.ucsd.edu/welkan/250B/cikmtutorial.pdf accessed 28 Jun 2017.
39. Emele MC, Dorna M (1998) Ambiguity preserving machine translation using packed representations. In proceedings of the 36th annual meeting of the Association for Computational Linguistics and 17th international conference on computational linguistics-volume 1 (pp. 365-371). Association for Computational Linguistics
40. Europarl: A Parallel Corpus for Statistical Machine Translation (2005) *Philipp Koehn*, MT Summit 2005
41. Fan Y, Tian F, Xia Y, Qin T, Li XY, Liu TY (2020) Searching better architectures for neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28:1574–1585
42. Fang H, Lu W, Wu F, Zhang Y, Shang X, Shao J, Zhuang Y (2015) Topic aspect-oriented summarization via group selection. Neurocomputing 149:1613–1619
43. Fattah MA, Ren F (2009) GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput Speech Lang 23(1):126–144

44. Feldman S (1999) NLP meets the jabberwocky: natural language processing in information retrieval. Online-Weston Then Wilton 23:62–73
45. Friedman C, Cimino JJ, Johnson SB (1993) A conceptual model for clinical radiology reports. In proceedings of the annual symposium on computer application in medical care (p. 829). Am Med Inform Assoc
46. Gao T, Dontcheva M, Adar E, Liu Z, Karahalios K DataTone: managing ambiguity in natural language interfaces for data visualization, UIST '15: proceedings of the 28th annual ACM symposium on User Interface Software & Technology, November 2015, 489–500, https://doi.org/10.1145/2807442.2807478
47. Ghosh S, Vinyals O, Strope B, Roy S, Dean T, Heck L (2016) Contextual lstm (clstm) models for large scale nlp tasks. arXiv preprint arXiv:1602.06291
48. Glasgow B, Mandell A, Binney D, Ghemri L, Fisher D (1998) MITA: an information-extraction approach to the analysis of free-form text in life insurance applications. AI Mag 19(1):59
49. Goldberg Y (2017) Neural network methods for natural language processing. Synthesis lectures on human language technologies 10(1):1–309
50. Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (pp. 19-25). ACM
51. Green Jr, BF, Wolf AK, Chomsky C, Laughery K (1961) Baseball: an automatic question-answerer. In papers presented at the may 9-11, 1961, western joint IRE-AIEE-ACM computer conference (pp. 219-224). ACM
52. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2016) LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems 28(10):2222–2232
53. Grishman R, Sager N, Raze C, Bookchin B (1973) The linguistic string parser. In proceedings of the June 4-8, 1973, national computer conference and exposition (pp. 427-434). ACM
54. Hayes PJ (1992) Intelligent high-volume text processing using shallow, domain-specific techniques. Text-based intelligent systems: current research and practice in information extraction and retrieval, 227-242.
55. Hendrix GG, Sacerdoti ED, Sagalowicz D, Slocum J (1978) Developing a natural language interface to complex data. ACM Transactions on Database Systems (TODS) 3(2):105–147
56. "Here's Why Natural Language Processing is the Future of BI (2017) " SmartData Collective. N.p., n.d. Web. 19
57. Hirschman L, Grishman R, Sager N (1976) From text to structured information: automatic processing of medical reports. In proceedings of the June 7-10, 1976, national computer conference and exposition (pp. 267-275). ACM
58. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
59. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991
60. Hutchins WJ (1986) Machine translation: past, present, future (p. 66). Ellis Horwood, Chichester
61. Jurafsky D, Martin J (2008) H. Speech and language processing. 2nd edn. Prentice-Hall, Englewood Cliffs, NJ
62. Kamp H, Reyle U (1993) Tense and aspect. In from discourse to logic (pp. 483-689). Springer Netherlands
63. Kang Y, Cai Z, Tan CW, Huang Q, Liu H (2020) Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics 7(2):139–172
64. Kim Y. (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882
65. Knight K, Langkilde I (2000) Preserving ambiguities in generation via automata intersection. In AAAI/IAAI (pp. 697-702)
66. Lass R (1998) Phonology: An Introduction to Basic Concepts. Cambridge, UK; New York; Melbourne, Australia: Cambridge University Press. p. 1. ISBN 978–0–521-23728-4. Retrieved 8 January 2011Paperback ISBN 0–521–28183-0
67. Lewis DD (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4–15). Springer, Berlin Heidelberg
68. Liddy ED (2001). Natural language processing
69. Lopez MM, Kalita J (2017) Deep learning applied to NLP. arXiv preprint arXiv:1703.03091
70. Luong MT, Sutskever I, Le Q V, Vinyals O, Zaremba W (2014) Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*
71. Lyman M, Sager N, Friedman C, Chi E (1985) Computer-structured narrative in ambulatory care: its use in longitudinal review of clinical data. In proceedings of the annual symposium on computer application in medical care (p. 82). Am Med Inform Assoc
72. Lyman M, Sager N, Chi EC, Tick LJ, Nhan NT, Su Y, ..., Scherrer, J. (1989) Medical Language Processing for Knowledge Representation and Retrievals. In Proceedings. Symposium on Computer Applications in Medical Care (pp. 548–553). Am Med Inform Assoc

73. Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies (pp. 142-150)
74. Mani I, Maybury MT (eds) (1999) Advances in automatic text summarization, vol 293. MIT press, Cambridge, MA
75. Manning CD, Schütze H (1999) Foundations of statistical natural language processing, vol 999. MIT press, Cambridge
76. Marcus MP, Marcinkiewicz MA, Santorini B (1993) Building a large annotated corpus of english: the penn treebank. Comput Linguist 19(2):313–330
77. McCallum A, Nigam K (1998) A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48)
78. McCray AT (1991) Natural language processing for intelligent information retrieval. In Engineering in Medicine and Biology Society, 1991. Vol. 13: 1991., Proceedings of the Annual International Conference of the IEEE (pp. 1160–1161). IEEE
79. McCray AT (1991) Extending a natural language parser with UMLS knowledge. In proceedings of the annual symposium on computer application in medical care (p. 194). Am Med Inform Assoc
80. McCray AT, Nelson SJ (1995) The representation of meaning in the UMLS. Methods Inf Med 34(1–2): 193–201
81. McCray AT, Razi A (1994) The UMLS knowledge source server. Medinfo MedInfo 8:144–147
82. McCray AT, Srinivasan S, Browne AC (1994) Lexical methods for managing variation in biomedical terminologies. In proceedings of the annual symposium on computer application in medical care (p. 235). Am Med Inform Assoc
83. McDonald R, Crammer K, Pereira F (2005) Flexible text segmentation with structured multilabel classification. In proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 987-994). Assoc Comput Linguist
84. McGray AT, Sponsler JL, Brylawski B, Browne AC (1987) The role of lexical knowledge in biomedical text understanding. In proceedings of the annual symposium on computer application in medical care (p. 103). Am Med Inform Assoc
85. McKeown KR (1985) Text generation. Cambridge University Press, Cambridge
86. Merity S, Keskar NS, Socher R (2018) An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*
87. Mikolov T, Chen K, Corrado G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems
88. Morel-Guillemaz AM, Baud RH, Scherrer JR (1990) Proximity processing of medical text. In medical informatics Europe'90 (pp. 625–630). Springer, Berlin Heidelberg
89. Morin E (1999) Automatic acquisition of semantic relations between terms from technical corpora. In proc. of the fifth international congress on terminology and knowledge engineering-TKE'99
90. Müller M, Salathé M, Kummervold PE (2020) Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503
91. "Natural Language Processing (2017) " Natural Language Processing RSS. N.p., n.d. Web. 25
92. "Natural Language Processing" (2017) Natural Language Processing RSS. N.p., n.d. Web. 23
93. Newatia R (2019) https://medium.com/saarthi-ai/sentence-classification-using-convolutional-neural-networks-ddad72c7048c. Accessed 15 Dec 2021
94. Nhàn NT, Sager N, Lyman M, Tick LJ, Borst F, Su Y (1989) A medical language processor for two indo-European languages. In proceedings. Symposium on computer applications in medical care (pp. 554-558). Am Med Inform Assoc
95. Nießen S, Och FJ, Leusch G, Ney H (2000) An evaluation tool for machine translation: fast evaluation for MT research. In LREC
96. Ochoa, A. (2016). Meet the Pilot: Smart Earpiece Language Translator. https://www.indiegogo.com/projects/meet-the-pilot-smart-earpiece-language-translator-headphones-travel. Accessed April 10, 2017
97. Ogallo, W., & Kanter, A. S. (2017). Using natural language processing and network analysis to develop a conceptual framework for medication therapy management research. https://www.ncbi.nlm.nih.gov/pubmed/28269895?dopt=Abstract. Accessed April 10, 2017
98. Otter DW, Medina JR, Kalita JK (2020) A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems 32(2):604–624
99. Ouyang Y, Li W, Li S, Lu Q (2011) Applying regression models to query-focused multi-document summarization. Inf Process Manag 47(2):227–237
100. Palmer M, Gildea D, Kingsbury P (2005) The proposition bank: an annotated corpus of semantic roles. Computational linguistics 31(1):71–106

101. Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Assoc Comput Linguist
102. Peng Y, Chi J (2019) Unsupervised cross-media retrieval using domain adaptation with scene graph. IEEE Transactions on Circuits and Systems for Video Technology 30(11):4368–4379
103. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
104. Rae JW, Potapenko A, Jayakumar SM, Lillicrap TP, (2019) Compressive transformers for long-range sequence modelling. arXiv preprint arXiv:1911.05507
105. Ranjan P, Basu HVSSA (2003) Part of speech tagging and local word grouping techniques for natural language parsing in Hindi. In Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003)
106. Rassinoux AM, Baud RH, Scherrer JR (1992) Conceptual graphs model extension for knowledge representation of medical texts. MEDINFO 92:1368–1374
107. Rassinoux AM, Michel PA, Juge C, Baud R, Scherrer JR (1994) Natural language processing of medical texts within the HELIOS environment. Comput Methods Prog Biomed 45:S79–S96
108. Rassinoux AM, Juge C, Michel PA, Baud RH, Lemaitre D, Jean FC, Scherrer JR (1995) Analysis of medical jargon: The RECIT system. In Conference on Artificial Intelligence in Medicine in Europe (pp. 42–52). Springer, Berlin Heidelberg
109. Rennie J (2000) ifile: An application of machine learning to e-mail filtering. In Proc. KDD 2000 Workshop on text mining, Boston, MA
110. Riedhammer K, Favre B, Hakkani-Tür D (2010) Long story short–global unsupervised models for keyphrase based meeting summarization. Speech Comm 52(10):801–815
111. Ritter A, Clark S, Etzioni O (2011) Named entity recognition in tweets: an experimental study. In proceedings of the conference on empirical methods in natural language processing (pp. 1524-1534). Assoc Comput Linguist
112. Rospocher M, van Erp M, Vossen P, Fokkens A, Aldabe I, Rigau G, Soroa A, Ploeger T, Bogaard T(2016) Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web, In Press
113. Sager N, Lyman M, Tick LJ, Borst F, Nhan NT, Revillard C, … Scherrer JR (1989) Adapting a medical language processor from English to French. Medinfo 89:795–799
114. Sager N, Lyman M, Nhan NT, Tick LJ (1995) Medical language processing: applications to patient data representation and automatic encoding. Methods Inf Med 34(1–2):140–146
115. Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A Bayesian approach to filtering junk e-mail. In learning for text categorization: papers from the 1998 workshop (Vol. 62, pp. 98-105)
116. Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P (2001) Stacking classifiers for anti-spam filtering of e-mail. arXiv preprint cs/0106040
117. Sakkis G, Androutsopoulos I, Paliouras G et al (2003) A memory-based approach to anti-spam filtering for mailing lists. Inf Retr 6:49–73. https://doi.org/10.1023/A:1022948414856
118. Santoro A, Faulkner R, Raposo D, Rae J, Chrzanowski M, Weber T, ..., Lillicrap T (2018) Relational recurrent neural networks. Adv Neural Inf Proces Syst, 31
119. Scherrer JR, Revillard C, Borst F, Berthoud M, Lovis C (1994) Medical office automation integrated into the distributed architecture of a hospital information system. Methods Inf Med 33(2):174–179
120. Seal D, Roy UK, Basak R (2020) Sentence-level emotion detection from text based on semantic rules. In: Tuba M, Akashe S, Joshi A (eds) Information and communication Technology for Sustainable Development. Advances in intelligent Systems and computing, vol 933. Springer, Singapore. https://doi.org/10.1007/978-981-13-7166-0_42
121. Sentiraama Corpus by Gangula Rama Rohit Reddy, Radhika Mamidi. Language Technologies Research Centre, KCIS, IIIT Hyderabad (n.d.) ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/
122. Sha F, Pereira F (2003) Shallow parsing with conditional random fields. In proceedings of the 2003 conference of the north American chapter of the Association for Computational Linguistics on human language technology-volume 1 (pp. 134-141). Assoc Comput Linguist
123. Sharifirad S, Matwin S, (2019) When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. arXiv preprint arXiv:1902.10584
124. Sharma S, Srinivas PYKL, Balabantaray RC (2016) Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script. In Proceedings of Language Resources and Evaluation Conference 2016 (pp. 47–51)
125. Shemtov H (1997) Ambiguity management in natural language generation. Stanford University
126. Small SL, Cortell GW, Tanenhaus MK (1988) Lexical Ambiguity Resolutions. Morgan Kauffman, San Mateo, CA

127. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642)
128. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26(1):320–322
129. Srihari S (2010) Machine Learning: Generative and Discriminative Models. http://www.cedar.buffalo.edu/wsrihari/CSE574/Discriminative-Generative.pdf. accessed 31 May 2017.]
130. Sun X, Morency LP, Okanohara D, Tsujii JI (2008) Modeling latent-dynamic in shallow parsing: a latent conditional model with improved inference. In proceedings of the 22nd international conference on computational linguistics-volume 1 (pp. 841-848). Assoc Comput Linguist
131. Sundheim BM, Chinchor NA (1993) Survey of the message understanding conferences. In proceedings of the workshop on human language technology (pp. 56-60). Assoc Comput Linguist
132. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems
133. Sworna ZT, Mousavi Z, Babar MA (2022) NLP methods in host-based intrusion detection Systems: A systematic review and future directions. arXiv preprint arXiv:2201.08066
134. Systems RAVN (2017) "RAVN Systems Launch the ACE Powered GDPR Robot - Artificial Intelligence to Expedite GDPR Compliance." Stock Market. PR Newswire, n.d. Web. 19
135. Tan KL, Lee CP, Anbananthen KSM, Lim KM (2022) RoBERTa-LSTM: A hybrid model for sentiment analysis with transformers and recurrent neural network. *IEEE Access, RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network*
136. Tapaswi N, Jain S (2012) Treebank based deep grammar acquisition and part-of-speech tagging for Sanskrit sentences. In software engineering (CONSEG), 2012 CSI sixth international conference on (pp. 1-4). IEEE
137. Thomas C (2019) https://towardsdatascience.com/recurrent-neural-networks-and-natural-language-processing-73af640c2aa1. Accessed 15 Dec 2021
138. Tillmann C, Vogel S, Ney H, Zubiaga A, Sawaf H (1997) Accelerated DP based search for statistical translation. In Eurospeech
139. Umber A, Bajwa I (2011) “Minimizing ambiguity in natural language software requirements specification, ” in Sixth Int Conf Digit Inf Manag, pp. 102–107
140. "Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research (2017) " AMIA ... Annual Symposium proceedings. AMIA Symposium. U.S. National Library of Medicine, n.d. Web. 19
141. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, (2017) Attention is all you need. In advances in neural information processing systems (pp. 5998-6008)
142. Wahlster W, Kobsa A (1989) User models in dialog systems. In user models in dialog systems (pp. 4–34). Springer Berlin Heidelberg, User Models in Dialog Systems
143. Walton D (1996) A pragmatic synthesis. In: fallacies arising from ambiguity. Applied logic series, vol 1. Springer, Dordrecht)
144. Wan X (2008) Using only cross-document relationships for both generic and topic-focused multi-document summarizations. Inf Retr 11(1):25–49
145. Wang W, Gang J, 2018 Application of convolutional neural network in natural language processing. In 2018 international conference on information Systems and computer aided education (ICISCAE) (pp. 64-70). IEEE
146. Wang D, Zhu S, Li T, Gong Y (2009) Multi-document summarization using sentence-based topic models. In proceedings of the ACL-IJCNLP 2009 conference short papers (pp. 297-300). Assoc Comput Linguist
147. Wang D, Zhu S, Li T, Chi Y, Gong Y (2011) Integrating document clustering and multidocument summarization. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(3):14–26
148. Wang Z, Ng P, Ma X, Nallapati R, Xiang B (2019) Multi-passage bert: A globally normalized bert model for open-domain question answering. arXiv preprint arXiv:1908.08167
149. Wen Z, Peng Y (2020) Multi-level knowledge injecting for visual commonsense reasoning. IEEE Transactions on Circuits and Systems for Video Technology 31(3):1042–1054
150. Wiese G, Weissenborn D, Neves M (2017) Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*
151. Wong A, Plasek JM, Montecalvo SP, Zhou L (2018) Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 38(8):822–841
152. Woods WA (1978) Semantics and quantification in natural language question answering. Adv Comput 17: 1–87

153. Xia T (2020) A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering Systems. IEEE Access 8:82653–82661. https://doi.org/10.1109/ACCESS.2020.2991328
154. Xie P, Xing E (2017) A constituent-centric neural architecture for reading comprehension. In *proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 1405-1414)
155. Yan X, Ye Y, Mao Y, Yu H (2019) Shared-private information bottleneck method for cross-modal clustering. IEEE Access 7:36045–36056
156. Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In data mining, 2003. ICDM 2003. Third IEEE international conference on (pp. 427-434). IEEE
157. Young SJ, Chase LL (1998) Speech recognition evaluation: a review of the US CSR and LVCSR programmes. Comput Speech Lang 12(4):263–279
158. Yu S, et al. (2018) "A multi-stage memory augmented neural network for machine reading comprehension." Proceedings of the workshop on machine reading for question answering
159. Zajic DM, Dorr BJ, Lin J (2008) Single-document and multi-document summarization techniques for email threads using sentence compression. Inf Process Manag 44(4):1600–1610
160. Zeroual I, Lakhouaja A, Belahbib R (2017) Towards a standard part of speech tagset for the Arabic language. J King Saud Univ Comput Inf Sci 29(2):171–178