CS F407 – Artificial Intelligence

Assignment - I

Natural Language Processing and its Applications

Name: Siddarth Suresh Gopalakrishnan

ID: 2017B3A71379H

Section: 1

Serial Number: 30

TABLE OF CONTENTS

INTRODUCTION

Natural Language Processing (NLP) is one of the fields of Artificial Intelligence which primarily gained importance due to its need in the field of linguistics and the study of language constructs for semantic analysis. It analyses how human beings understand and use languages, and tries to impart these skills into a computing system by developing various tools and algorithms. NLP had its roots in the 1950s when Alan Turing proposed the Turing test as a rubric of intelligence of computing systems, which requires automated recognition and generation of natural language. NLP now owes its foundation to many sciences such as mathematics, linguistics, computer sciences and information systems, electrical and electronics engineering, psychology, robotics and automation, etc.

Linguistic sciences help us explain the linguistic observations around us, such as in communications, media, writing, etc. A significant part of linguistics is to understand how humans perceive this information, how they communicate or respond, and understand the spoken language in general. It also delves in understanding the meaning of the utterances and how it relates to the current environment. A part of this field also goes into understanding the constructs of the language and the way its words, phrases, and sentences are structured.

In the 1950s, the hypothesis of symbolic NLP was summarised by an American philosopher J. Searle, where the computing system understood natural language by applying a collection of rules on the input data. Till around the 1980s, natural language processing was emulated by these collections of pre-written rules. However, the late 1980s marked the advent of statistical NLP where the introduction of various machine learning techniques and increased computational resources, improved the performance of different NLP systems and helped in automating the entire process, which thus required lesser human supervision. In modern days, due to the increasing popularity in deep-learning techniques and automated representation learning, the process of implementing NLP models became much more straightforward. This led to improved accuracy of the models, which produced state-of-the-art results in many NLP tasks, thus marking the era of Neural NLP.

The structure of linguistic models involves the following key components:

- Phonology – This is the study of the sounds and their patterns for various languages. It also involves organizing different speech sounds in order to convey the desired meaning. Phonology is related to various fields of linguistics, such as cognitive science, language acquisition, etc.
- Lexical – This process splits the input data or sentence into fragments or words and extracts the meaning of each word.
- Syntax and Semantics – This process analyses the tokens generated and checks if these tokens follow the grammatical format and obey the rules of the language.
- Morphology – This analyses how the words of a language are formed.

- Pragmatics – One of the crucial aspects of NLP, which tries to analyze and understand the meaning of the sentence containing ambiguity, underlying emotions, connotations, etc.

The crucial challenge in the field of natural language processing is making sense out of the sequential and unstructured data, data which doesn't fit neatly in a traditional tabular format. Human language is very complex to interpret as it is very expressive and symbolic, meaning there are multiple ways of conveying the same meaning to the other person. There are also many ways of arranging words in a sentence which results in multiple meanings, making this a difficult subject. This domain has come a long way from just trying to understand how to interpret a text or speech based on the keywords present, to now trying to figure out the meaning and connotations behind those words.

This discipline focusses on the interaction between the human language and the processes, systems and algorithms to extract useful information and gain insights. Due to the increase in the amount of available data and the technological improvements in computational facilities, it is easier for researchers and data scientists to achieve meaningful results due to which this field is finding interesting applications in lots of industries, such as business, marketing, governance, etc. But, there are still some challenges faced by researchers in this domain, termed as AI-Hard problems, which are broadly categorized into natural language understanding, speech recognition and natural language generation.

NLP is also helpful in past-of-speech tagging (POS tagging) application which is the process of categorizing words into their respective parts of speech or lexical categories. Part of speech categories include nouns, verbs, adjectives, pronouns, etc. This task requires a lot of experience to effectively categorize words as a normal bag-of-words approach might not work always. Rule-based POS tagging make use of a dictionary to tag each word of a sentence or phrase. The ambiguity in the bag-of-words method can be removed in the rule-based method as it can understand the context of the current word by analyzing words before and after it to estimate its part of speech correctly. The probabilistic approach of POS tagging has two ways of going about this task, namely the word frequency approach, and the other is the tag sequence probability approach. The former method calculates the probability of a word attaining a particular tag. This means that we first calculate the tag which a word takes most frequently and assign that tag to the same word in ambiguous cases. The latter approach calculates the tag of a word based on the probability from the tags of all the words occurring in the sequence.

Natural language understanding is a sub-domain which primarily deals with discovering the meaning of input texts and speeches, which basically tests the computer's ability to read and comprehend the input content. This closely ties to the field of human-computer interaction (HCI) where the aim is for the computing system to understand human commands without any formalized structure or syntax and communicate back intelligently. This is a challenging task in the field of computing, as the computer has to understand not only the input sentence but also the intent behind the command, while also accommodating human errors such as mispronunciations and other syntactical errors. There are consistent efforts in

researching more about this field which leads to the creation of chatbots which can have intelligent conversations with the user without much supervision. Prime examples of such inventions are Siri by Apple, the Google Assistant by Google, Cortana by Microsoft, Alexa by Amazon etc.

Speech Recognition deals with enabling the computer to interpret words spoken by humans and translate it into text format. This field also has been increasing in popularity along with voice recognition systems, whose purpose is to identify the speaker from their pattern of speech, their accent and various other factors. Speech and voice recognition are sometimes used interchangeably, but they mean different things. There is a wide range of application in this field, ranging from household appliance control, dictation solutions, mobile devices and is also scalable in various industries such as automobile industries, healthcare, education, etc. There are however some areas which need more research and improvement such as accounting for human errors and variations in their pronunciation, random noise, pitch, accent, etc. which could lead to inaccuracies in recognizing the word, sentence, or user.

Natural language generation is the process where the computing system takes in some structured data as the input and produces meaningful sentences or phrases which explain the input data in a human-understandable form at very high speeds. Natural language understanding reads the unstructured input human language, whereas natural language generation only outputs natural language that explains its structured input data. This is one of the fastest-growing technologies in the field of NLP and is quickly being scaled to many industrial applications. The main reason for its increasing popularity is that people wanted computing systems to communicate ideas in a human-understandable format from the available structured data at high speed and accuracy. Due to the expanding research in this field and the increased computational powers, this field has a large number of applications such as generating text summaries, data analytics, auto-generated reports, chatbots, speech translation, etc. But, with the growing applications, there are also various challenges associated with it. Human language is, in general, very complex and has a lot of ambiguity, variation, and hidden meaning due to the high expressivity. This still leaves some gap for optimization in order to improve the accuracy of understanding the human language and commands, of the generated text, of correctly identifying the user and their voice patterns, etc. thus bridging the distance between the users and the computing systems.

One of the primary applications of NLP is Machine Translation. As the name suggests, it is the task of converting one natural language to another, while still keeping the meaning of the original sentence and producing syntactically correct output. There is still a lot of research to be done in this domain to improve accuracy in order to maintain proper grammar, syntax, meaning, format, etc. Even nowadays, we can sometimes observe that Google Translate may alter the meaning of the output sentence or not structure the output in a proper understandable manner. Closely tied to the aforementioned use-cases and applications is Sentiment Analysis, which is also a successful application of NLP. Sentiment analysis deals with understanding the sentiment of a sentence or text. It also tries to identify the emotion behind the input statement, which isn’t expressed explicitly. It has a variety of applications such as computational

linguistics, understanding customer reviews or ratings, collecting opinions, etc. These not only give a better idea of what the general customer demographic likes and dislikes but also helps the company to evaluate the quality of the services it provides.

Spam filtering is also a popular application of NLP, one which most beginners start their NLP career with. This is the process of filtering out unwanted emails or emails from unknown or potentially harmful sources. NLP can be used for spam filtering by analyzing the false-positive and false-negative cases. We could use n-gram models or even Bayesian classification for this task. The latter is more widely used nowadays as we conduct an analysis of the occurrence of a word with respect to its occurrence in a large corpus of words which typically appear in spam emails. We could also analyze semantic similarity, which is helpful in addressing the problem of spam detection based on similar keywords present in the content.

Given so many applications and advantages of NLP, it may seem like the research in this domain is sufficient, or almost all the application fields have been exploited, but NLP does have its limitations and disadvantages. One primary disadvantage of NLP is to decode complex statements or ambiguous statements with hidden meanings. NLP is not as developed to fully understand sarcasm, or irony or even jokes. The ability to dynamically attaining knowledge about a conversation is still lacking. What the person in front means actually, the hidden meaning, underlying connotations, specificity, etc., aren't fully understood. As stated earlier, natural language processing still hasn't perfected the task of machine translation. This is clearly evident as even Google Translate sometimes doesn't guarantee a good translation. The meanings may be altered, or the structure and framing of the sentence could be incoherent. But, given the amount of data nowadays, and also the immediate need to process them, be it for businesses, or customer review understanding, the job becomes increasingly difficult for humans to do; hence, there is an increasing need of automated assistance.

With the above use-cases and information, it is clear that Natural Language Processing will be the future. With the increase in the amount of available text data, there's an increasing need for well-developed NLP models to understand the data accurately and quickly. Many industries already use the powers available because of NLP for enhancing the customer experience with the help of chatbots, personal assistants, information extraction. There are also some popular applications such as email spam detection, translation, report summarization, various medical applications, etc. Currently, NLP models can understand just the input text or speech and answer appropriately, but, with the consistent research, there is scope for these intelligent systems even to understand the connotations behind the words efficiently and also interpret the emotions behind the sentence. NLP has changed the way humans interact with computers, and these technologies will drive the change from data to an intelligence-driven era in the years to come.

LITERATURE SURVEY

The following section discusses recent trends and research conducted in the field of NLP and also reflects on some of the future goals and prospects of this domain. It discusses various applications and even some language or sequence models best suited for that application.

Michael Jordan and Tom Mitchell (Jordan, 2015) discuss the future prospects and trends in the field of machine learning and share their aim of computer programs continuously gathering experience and improve their performance. Machine learning and statistics are the core subjects of Artificial Intelligence and Data Science, which is growing in popularity due to the increase in the amount of available data, and improvement in the technology of computing systems. There are a lot of algorithmic paradigms developed for the computing systems to learn from and tackle a wide variety of problems, some of which are supervised learning, unsupervised learning, reinforcement learning, etc. It is, however, essential to note that no one algorithm works for all the tasks. Although these paradigms help in organizing ideas and thoughts, the research nowadays requires a blend of all the categories, such as semi-supervised learning.

Erik Cambria and Bebo White (Cambria, 2014) talk about the evolution of NLP reflecting on the semantics, syntactical analysis and pragmatics. There is a lot of data available on the internet, and a lot of algorithms which can efficiently extract texts from the internet, but there are very few algorithms which understand the meaning of those texts. But with the increasing amount of data, our models need to be fast and efficient. Syntactics analyses the individual words of a text, but with the growing amount of user-generated content, keyword analysis won't work because of increasing spam comments. Due to this, NLP systems should jump to analyzing the content of the input rather than just keywords, which is the semantic analysis. Syntax and semantics from the common sense of a sentence which enables us to extract more information. This, however, doesn't wholly capture the more abstract details such as context, or hidden meanings. For this, we must design our models to be more adaptive by analyzing the pragmatics.

PM Nadkarni (Nadkarni, 2011) discusses the various fundamental models developed in handling NLP tasks since the advent of the statistical NLP era (the mid-1990s). The reason for the failure of the symbolic or rule-based period was its highly restrictive nature, due to which a lot of rules had to be handwritten. As the amount of data increased, it became difficult to manage so many rules, which is when statistical methods showed elegance by learning from the vast amount of data. It became easier to handle low-level activities of NLP such as POS tagging, tokenization, morphological decomposition, etc. As the number of features in a generic NLP task can be relatively high, discriminative models would work better than generative models. Nadkarni (Nadkarni, 2011) and Diksha Khurana (Khurana, 2017) discuss various models for general NLP tasks. SVMs can be used for segregating words into parts of speech using a demarcating hyperplane using a kernel function. Hidden Markov Models

(HMM) is a probabilistic model which can give outputs depending on the states with associated probabilities. HMMs are thus widely used for speech recognition tasks where the output is matched to an input sequence of phonemes which are more likely to produce that output. Conditional Random Fields (CRFs) are also discriminative models which are similar to HMMs in a sense; it generalizes Logistic regression to sequential as HMM generalizes Naive Bayes.

Tom Young and D. Hazarika (Young, 2018) talk about the development of various deep learning models which have paved the way for state-of-the-art results in NLP domain. For many years, NLP models were being designed with respect to shallow models such as logistic regression or SVMs. Now, due to the success of deep learning techniques in representation learning, and word embeddings, NLP problems are approached with neural networks and dense vector representations enabling even a simple deep network to perform tasks such as Named-Entity Recognition (NER) and Part of speech (POS) tagging, easily. Nowadays, with the advent of reinforcement learning and deep generative models, NLP tasks have become easier. The author also discusses the results of implementing these deep learning on some datasets, e.g., implementing a simple bidirectional LSTM model on the WSJ-PTB dataset having 1.17 million tokens, resulted in a per-token accuracy of 97-98%.

Roman Collobert (Collobert, 2008) discusses multi-task learning in NLP. NLP problems can be broken down into sub-problems, similar to the divide and conquer approach, which can then be performed jointly using weight sharing. Such sub-tasks involve, POS tagging, analyzing semantics, chunking etc. The more complex the task at hand, the more is the necessity for a unified architecture which can execute the sub-tasks together for more efficiency. In the modern era, this is possible with the help of deep neural networks, which can learn the representations and features with low initial knowledge. For this, a lookup table is generated where each word of a finite dictionary is embedded. Each word of the input sentence is then converted to vectors (word-vector representation). After taking care of other factors such as word representation variability, length of the sentences, etc. we pass our representations to the convolutional models or time-delay neural networks for long term dependencies. Multi-tasking is then achieved by sharing lookup tables which improves the training process, which improves performance.

Wang et al. (Wang, 2018) talk about multi-task analysis platform, GLUE, for natural language understanding. GLUE stands for General Language Understanding Evaluation which is a benchmark, which is a collection of datasets used for training and evaluating NLP models with respect to various tasks and also relative to other models. This benchmark is necessary in order to establish a certain standard of performance for various NLP activities. The benchmark performance is determined by the performance of current transfer learning and reinforcement learning methods. One proposed pre-training technique was ELMo (Embedded from Language Models), which is one of the best multi-tasking models. On the basis of this metric, it was found that multi-task models or other transfer learning techniques, in general, perform better than individual models per task.

Word representation is the process of converting raw data into a vector or other suitable representation for easier understanding by the machine learning model. Mikolov et al. (Mikolov, 2013) explains the distributed representation of words and also provide improvements to the word2vec skip-gram model. The skip-gram model is a context predicting model which is useful for getting word representation for predicting the neighbouring words. The author discusses two improvements, namely, noise contrastive estimation and negative sampling. The difference between them is that the latter doesn't calculate probabilities for noise distribution and also uses sub-sampling methods which is a better approach for representing uncommon words. The authors also suggest deriving phrase representations to understand and recognize phrases or idioms.

As discussed previously, the ELMo (Embedded from Language Models) model is considered one of the best multi-tasking models. Peters et al. (Peters, 2018) talk about ELMo, which is a deeply contextualized model. This technique has the ability to not only model word semantics and syntax, but also word polysemy, which is the situation when words or phrases take multiple meanings. ELMo is a bidirectional LSTM model which is coupled with Language Models and trained on a large text corpus. ELMo representations are a function of the entire input sequence, unlike other embeddings. The vector representations from this model, when added to the existing models improve the state-of-the-art of many applications such as sentiment analysis, question answering, etc. On experimenting the effectiveness on various benchmark activities such as Question Answering (SQuAD), Named Entity Recognition (CoNLL 2003 dataset), it was found that ELMo, on an average, performs 2-3% better than the baseline models.

Ehud Reiter (Reiter, 1996) talks about building natural language generator systems and the way a basic NLG system functions. Natural language generations are the process of generating human-understandable output after understanding structured input data. The initial phase for an NLG system is content determination, which is to determine what information must be mentioned. After this, the structuring of the data must be taken care of, in order to convey information coherently. This leads to the phase of sentence planning, where the sentence must be cohesion among sentences, is established using grammar tools such as pronouns, conjunctions, etc. Sentence planning is one of the most important phases of NLG as it makes the output more readable and fluent. The final stage of an NLG system is Realization which is the process of actually creating the sentence. This output sentence must adhere to the syntax of the language, morphology and orthography, which comprises of capitalization, punctuation, spellings, etc.

Schuster and Paliwal (Schuster, 1997) elaborate on BRNNs (Bidirectional Recurrent Neural Networks) and their impact on NLP tasks. BRNN is just an extended form of RNN with the past also depending on the future for semantic coherence. Aside from the regular forward pass, there is also a backward pass, where each output word is predicted based on two activations instead of one. The author experimented on the TIMIT speech database for comparing the performance of various ANN models in a task of classification of phonemes. The TIMIT database contains 6300 sentences by 630 speakers with training, and testing data

split such that 142910 phoneme segments were for training and the rest for testing. The result was that the uni-directional RNNs (forward and backward) gave a recognition rate of only around 65%, whereas the BRNN model gave a recognition rate of approximately 70%.

Chung et al. (Chung, 2014) discuss variants of RNNs such as LSTMs and Gated Recurrent Units (GRU). LSTMs generally work well as sequence models which preserve long-term dependencies, whereas GRUs are found to be better suited for machine translation tasks. GRUs have an additional memory cell and a gamma parameter which control when to hold and when to update the value of the memory cell. GRUs are similar to LSTMs with forget gate, but they require fewer parameters than the latter. There are some tasks such as speech signalling modelling and polyphonic music modelling where the GRU model performs similar, if not better than, LSTMs. But there are other tasks where the LSTMs clearly outperform GRUs such as unbounded counting, which GRUs can't perform. As a conclusion to some experiments which the author performed, it is observed that for relatively infrequent datasets and a fixed number of parameters, the GRUs outperform regular LSTMs.

Wolf et al. (Wolf, 2019) present a paper talking about transformers. One of the main areas of growth of NLP is the improved architecture and pre-training of the models. One of the striking advances with respect to improved architecture is the transformer, which is increasingly being used in NLG and NLU tasks. The main reason this architecture is gaining popularity is because of parallel training and that it also records long-term dependencies. Many researchers use this as a foundation to model more complex architectures which can easily handle more data efficiently. One popular model developed by Google was BERT, which is Bidirectional Encoder Representations from Transformers. BERT is especially useful in training deep bidirectional models from a large unlabelled corpus, which when incorporated with another output layer, will work effectively as a state-of-the-art model for tasks such as question-answering models, etc. without much change.

G. Lample et al. (Lample, 2016) talks about one of the most popular applications of NLP, which is, Named Entity Recognition (NER). NER requires a lot of domain-related knowledge as the models have to learn from a very small amount of supervised data. The author discusses two models best suited for this, bidirectional LSTM with CRF and Stack LSTM. LSTM (Long Short Term Memory) models are one of the most popular models which are used for long term dependency conditions, which isn't handled well by normal RNNs. Bidirectional LSTMs are used for backward dependencies as well, in coherence with the semantics. CRF considers future observations and also that the features are inter-dependent. These models are then used to derive the word embeddings which will also be helpful in transfer learning. The author discusses an example of how a bidirectional LSTM on CoNLL 2003 dataset, along with character and word embeddings, resulted in an accuracy of 91%.

Lai, Xu et al. (Lai, 2015) talks about models for the text classification task, which is the process of tagging sentences or segregating them into categories. The model considered for this paper is the recurrent CNN due to the development of word embeddings. The reason they're preferred over regular RNN is that RNNs capture the context of the sentence, but tend to prefer

recent inputs. Recurrent CNNs are BRNNs, which extract the important keywords with the help of word representations and max pooling, although learning the kernel size may be tedious. The models were then compared using some well-known datasets such as 20newsgroups, Stanford Sentiment Bank, ACL set and Fudan set. It was found that the recurrent CNN gave better results in all four datasets but was a bit slower in performance than the regular RNN for the time it takes to generate a sentiment tree.

Neural machine translation is a popular application of NLP and a neural network approach of the previous statistical machine translation process. It is a sequence-to-sequence model which uses an encoder and a decoder network. Cho, Bengio, et al. (Cho, 2014) discuss two models for this purpose, an RNN Encoder-Decoder and a Gated recursive CNN (grConv). Both models have different encoders, but the same Gated RNN decoder. In the encoder-decoder approach, the main objective is learning the conditional distributions for the output sentence. After this, one can apply some search algorithms such as Beam Search, for maximizing joint probability and not individual output word probability, as in the case of Greedy Search. The encoder then generates a vector representation, which the decoder uses to generate the translated output.

Luong et al. (Luong, 2015) discuss the use of an attention model instead of a regular Encoder-Decoder RNN. One of the main reasons for using an attention model is because a regular BRNN might not effectively capture the context for long sentences as it cannot store that much memory. Usually, for any machine translation task, people generally use some variation of RNNs, as discussed earlier. The main problem with RNNs is faced during the process of decoding such large representations. Due to this, the general context or meaning of the sentence might be lost. In order to tackle this, we introduce another hidden layer which takes in a context parameter and also keeps track of how much attention should be paid. The author also discusses an experiment conducted on WMT'14 English to German translation, with the model trained on the newstest2013 dataset, with a case-sensitive BLEU rubric. It was found that a general RNN architecture gave an average BLEU score of around 20 whereas a basic Attention Model gave a BLEU score of about 23.

Text Summarisation (Nenkova, 2012) is an essential application of NLP. Broadly there are two approaches, namely, the topic representation-based approach and the indicator-based approach. In the early 1950s, there was a lot of research done on word frequency which also included in identifying spam, or very high-frequency words. This is one of the driving concepts of the field of text summarisation. In the topic representation approach, each sentence is first analyzed to figure out how much it is related to the topic or to identify explanatory words. In the indicator representation approach, the sentences are represented on the basis of some features in order to rank their importance, without any special emphasis on the topic of the sentence. In each of the methods, the representations are then given some scores based on their relevance and importance, using some machine learning algorithms. The sentences are then selected Greedily in order to summarise the entire document or report.

Hirschman (Hirschman, 2001) talks about the impact of NLP in Question Answering systems. This is a field which is closely tied to NLP and IR, which, as the name suggests, answers questions asked to the system. Modern internet domains can provide a list of documents, ranked with respect to relevance, but it doesn't provide answers. There are various steps in designing a question-answer system. One must first make sure that the model is able to interpret the meaning of the question posed. One must then design the model not only to fetch the most likely answers but also enable the model to choose the best answer, which justifies the question. Choosing the best possible answers is also an ordeal as the model has to fetch appropriate answers depending on whether it's a factual answer, or an opinion answer or a summary. There are a lot of use-cases of this application ranging from search engines, to help systems, or even companion learning systems in the education domain.

Robert Moore (Moore, 1999) talks about a widespread application of NLP, speech recognition. In order to have efficient language models for speech recognition, we need to account for the long-term dependencies which sentences have while predicting the output. Due to this reason, short term n-gram models aren't sufficient to produce sensible outputs with high accuracies. The author discusses some architectures for natural language-based language models such as word lattice parsing and dynamic generation of partial grammar networks. The former is one of the oldest approaches where the recognizer generates specific output word hypotheses and assigns some scores to the potential start and endpoints corresponding to each word in the sentence. The language model then traverses the word lattice to find the path with the highest score.

Liu and Zhang (Liu, 2012) wrote a chapter on sentiment analysis and opinion mining which elaborates on the topic and discusses some of the challenges. Sentiment analysis is the process of extracting people's opinions towards certain instances, people, products, etc. It is the backbone of customer review technology which has a huge impact on the company as it will help them analyze a customer's preferences. It helps them figure out which areas to improve customer experience. This is a key technology as this task would be very difficult for a human to do and is also free from human bias. The general method is polarity classification which is a form of binary classification of a review to classify it as overall positive or negative opinion. Due to the high amounts of available data, it's first categorized using text classification algorithms and also grouped by features for a useful summary.

Hussein D.M (Hussein, 2018) performs an analysis on various other research papers about the challenges posed by sentiment analysis tasks. With the amount of data on the internet, we need efficient text mining techniques to extract useful meaning from them. The challenges faced may be classified broadly into theoretical and technical challenges. Some of the most common ones are spam filtering, bi-polar words, the intensity of opinion, a huge amount of lexicons, or even model related such as overheads in the execution of our NLP model. As a review has many components such as the object, the scale or intensity of the opinion, or even hidden meanings such as innuendos or sarcasm, etc., we must consider how these components work together to generate meaning. A straightforward keyword processing technique or text summarization technique will not solve this problem.

CONCLUSION

Natural Language Processing has changed the way we analyze, and communicate with, the outer environment. It is a science which uses various mathematical methods to analyze human language to help humans communicate more effectively with computing systems. NLP is the field which combines the laws of linguistics and natural language to advanced computing techniques such as machine learning and artificial intelligence. Although the research in this domain has helped immensely in the development of new intelligence agents, there is still a lot of development needed to improve the efficiency of existing systems and model new ones.

There is increasing use of NLP by online e-commerce companies to use NLP as a tool for targeted advertisements. By extracting data from social media, searches on the internet, customer behaviour, etc., NLP can understand the pattern and identify what items are more in demand by someone and thus, offer a personalized advertisement. Similarly, there are a lot of companies using bots for customer service agents which have developed to such an extent that they can have a full conversation about the users' problems and find the correct or most relevant solution.

The future of NLP has to focus more on customer service and hospitality. We can already see the slow rise of chatbots and voice assistants, which gets straight to the point to answer the user's queries immediately. This is a prime example of natural language understanding, text or speech recognition and efficient implementation of question-answering systems. As it is customer-oriented, chatbots must be designed to quickly understand the user and respond with the correct or relevant answer. But this only enables them to respond to short and straightforward requests. In order to understand more complex statements and longer requests, we must club our chatbot with cognitive deep learning models which can better understand the semantics and the meaning of the statement.

Given the increasing amount of data, it is essential to make NLP models more lightweight in order to improve efficiency with the given computational facilities. NLP models, in general, are known to train and fine-tune more parameters than the regular neural networks, due to which, the models might not be that efficient. For some complex tasks such as NLG or speech recognition, the number of parameters to train becomes too large, which increases with the increase in the text corpus or input data. There are many techniques which can be used for implementing faster and smaller models:

- Distillation – This process is based on transfer learning, where we take BERT as our base model while training our current model. We train our current model in such a way that it reproduces the generalization capabilities, e.g. DistilBERT whose performance was 95% that of regular BERT, which was 40% smaller and 60% faster.
- Pruning – In this method, we directly work on the reference or base model, and optimize or remove unnecessary components to reduce the size. There are various methods for pruning such as head pruning, weights pruning, layer pruning etc.

- Quantization – This reduces computation cost by rounding off floating-point numbers to integer weights.

Natural Language Processing, thus, has a very bright future with an immense number of applications. Like any other science field, there is a lot of research and innovation yet to be done in this domain and artificial intelligence in general. The primary focus on natural language processing as of now is to make the computing systems smarter in enabling them to understand better the semantics, the underlying meaning, the sarcasm, the polysemy of phrases and words, etc. The power of making unstructured data understandable to the computing system must be harnessed to the fullest.

REFERENCES

Cambria, E. &. (2014). A review of natural language processing research. *IEEE Computational intelligence magazine, 9(2)*, 48-57. Retrieved from https://scihub.wikicn.top/https://ieeexplore.ieee.org/document/6786458

Cho, K. V. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR*. Retrieved from http://arxiv.org/abs/1409.1259

Chung, J. G. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*. Retrieved from https://arxiv.org/pdf/1412.3555.pdf?ref=hackernoon.com

Collobert, R. &. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). Retrieved from https://dl.acm.org/doi/pdf/10.1145/1390156.1390177?casa_token=qveQ50jhL40AAAAA:bCdHKQn0sj9ekGarUPMuEKf6q8ZjxxNJ7fMOQyyXETUaC0yTTMixVnbS5PNTZd2tC31Ix83Il9f0cQ

Hirschman, L. &. (2001). Natural language question answering: the view from here. *Natural language engineering*, 275-300. Retrieved from https://www.researchgate.net/profile/Rob_Gaizauskas/publication/231807195_Natural_Language_Question_Answering_The_View_from_Here/links/0c96052a09fa7b819e000000/Natural-Language-Question-Answering-The-View-from-Here.pdf

Hussein, D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 330-338. Retrieved from https://www.sciencedirect.com/science/article/pii/S1018363916300071

Jordan, M. I. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 255-260. Retrieved from https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf

Kacprzyk, J. &. (2010). Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation. In *IEEE Transactions on Fuzzy Systems* (pp. 461-472). Retrieved from https://ieeexplore.ieee.org/document/5382512

Khurana, D. K. (2017). Natural Language Processing: State of The Art, Current Trends and. *CoRR*. Retrieved from http://arxiv.org/abs/1708.05148

Lai, S. X. (2015). Recurrent convolutional neural networks for text classification. *Twenty-ninth AAAI conference on artificial intelligence.* Retrieved from http://zhengyima.com/my/pdfs/Textrcnn.pdf

Lample, G. B. (2016). Neural architectures for named entity recognition. *CoRR*. Retrieved from http://arxiv.org/abs/1603.01360

Liu, B. &. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Boston, MA: Springer.

Luong, M. T. (2015). Effective approaches to attention-based neural machine translation. *CoRR*. Retrieved from http://arxiv.org/abs/1508.04025

Mikolov, T. S. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119. Retrieved from https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Moore, R. C. (1999). Using natural-language knowledge sources in speech recognition. In *Computational Models of Speech Pattern Processing* (pp. 304-327). Springer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.2954&rep=rep1&type=pdf

Nadkarni, P. M.-M. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*, 544-551. Retrieved from https://doi.org/10.1136/amiajnl-2011-000464

Nenkova, A. &. (2012). A survey of text summarization techniques. In *Mining text data* (pp. 43-76). Boston, MA: Springer. Retrieved from https://link.springer.com/chapter/10.1007%2F978-1-4614-3223-4_3

Peters, M. E. (2018). Deep contextualized word representations. *CoRR*. Retrieved from https://arxiv.org/pdf/1802.05365.pdf%E3%80%91

Reiter, E. &. (1996). Building natural language generation systems. *CoRR*. Retrieved from http://arxiv.org/abs/cmp-lg/9605002

Schuster, M. &. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing, 45*, 2673-2681. Retrieved from https://scihub.wikicn.top/https://ieeexplore.ieee.org/document/650093

Wang, A. S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*. Retrieved from https://arxiv.org/pdf/1804.07461.pdf

Wolf, T. D. (2019). Transformers: State-of-the-art natural language processing. *CoRR*. Retrieved from https://arxiv.org/pdf/1910.03771.pdf

Young, T. H. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 55-75. Retrieved from https://scihub.wikicn.top/10.1109/MCI.2018.2840738