Text Recognition and Generation using Natural Language Processing

Muhammad Hasnain Khan
Dept. Of Computer Science
FAST NUCES Islamabad
i191255@nu.edu.pk

***Abstract*—Natural Language Processing a way used in computers to understand the human language. Automatic summarization, speech recognition, topic segmentation, translation and sentiment analysis can be done using Natural Language Processing techniques. NLP can be categorized as one of the most difficult field in Artificial Intelligence because of nature of human language. There is a huge diversity in human language. Let's say there is a sentence as "He was found by mountain". Two meanings can be extracted from the sentence one is that mountain can be a name of someone and second is mountain is a noun. Human language is complex as it can be seen in the sentence. It is really difficult for computers to understand the grammar rules. But with the help of Natural Language Processing techniques we can train computers to understand the language and generate text according to our needs.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Natural Language Processing (NLP) is a sub-field of AI which is used to aid computers to understand the human's natural language. As a human, we speak and write in multiple languages. But a computer's native language known as machine code is difficult for humans to understand. NLP is the ability of the computer software to understand human language. Most NLP methods depend on AI to separate outcomes from tests through human language. An interaction between computers and humans can be broke down into these steps. i): Human Talks to the machine ii): Machine listen to audio iii): Conversion of audio to text iv): Different analysis on the text (Which will be explained below) v): Generation of the resultant text vi): Text to audio conversion vii): Machine respond to the human by playing the audio. Normal Language Processing is the major power behind some amazing applications like Google Translate, Word processors, OK Google, Siri, Cortana and Alexa. Nowadays NLP is considered one of the most difficult problem in AI because nature of human language makes it difficult. Computers are not able to understand the human language truly. Current approaches of NLP are based on deep learning, a sub-field of AI that examines the data at a deeper level. Deep learning models require a huge amount of structured data to train and test on. According to the 21st century, 22% of the data is in structured form. Data has been generated as a speech, tweet, messages etc.

A. Sentence Segmentation

The first step in the NLP is the sentence segmentation which break text into the separate sentences. There are some techniques where sentence split can be placed on the basis of punctuation mark but advance techniques use more complicated techniques that work even if a document is not well formatted. Sentence: I am Muhammad Hasnain Khan. I am a graduate student in FAST. After Sentence Segmentation we will have two sentences. 1: I am Muhammad Hasnain Khan. 2: I am a graduate student in FAST.

B. Word Tokenization

After splitting a sentence into the sub-sentences we can split those sub-sentence into words. After tokenization the first sentence will be tokenized into words as "I", "am", "Muhammad", etc.

C. Parts of Speech Tag

After step 2, we'll have a detailed look at the each token and will try to predict its part of speech, whether it's a verb, a noun, a preposition, etc.

D. Text Lemmatization

In many languages, words exist in multiple forms like singular or plural. Finding the singular or plural in NLP is called Text Lemmatization.

E. Named Entity Recognition (NER

The purpose of Named Entity Recognition is to detect and label found nouns with the real-world concepts that they represent. After NER, we will have a useful presentation of our sentence.

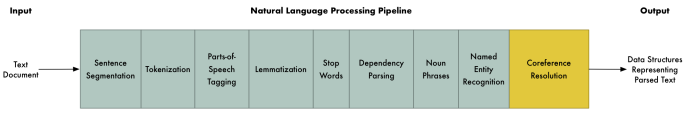

Fig. 1. NLP Pipeline.

Companies like Google and Yahoo classify and filter emails with the help of NLP though deep learning analysis of the text in emails. NLP is being used in talent recruitment by identifying the skills of potential hires. NLP is improving disease diagnosis and brining costs down while healthcare organization are going through a growing adoption of electronic health records.

II. Critical Review

A. Automated Genre Classification of Books Using Machine Learning and Natural Language Processing

It is a difficult task for a human to read the entire book classify the genre of the book. In this paper, size of the initial matrix is reduced using Wordnet Principle Component Analysis and then AdaBoost classifier is applied to predict the genres of the book. Unlabeled and named information is utilized to comprehend the structure of information and utilized before the preparation of an AI model. Named information is delivered from books whose type is known and unlabeled information is created from books whose kind isn't known. Named information is delivered from the books whose type and information is unlabeled. Weakness of this paper is that most of the things are not tackled well enough and they are assuming most of the things to be in the way so that AdaBoost will work with Wordnet. In the start of the paper, they mentioned that size of the initial matrix will be reduced to classify the genre of the books. The methodology mentioned in the paper does answer with the initial problem statement. Yes, the approach and results are critically analyzed with the help of tables and pictures. Yes, there are other possible interpretations as the proposed technique is for predicting the genre of the book but it can be scale up and can be used to predict article, magazines etc. Yes, all the technical details are completely explained and it makes sense that how NLP is used to convert the text from the books and how it can be used in machine learning model by reducing the size of the feature matrix. Yes, the results can be verified by doing Wordnet, cleansing data, preprocessing and applying machine learning model in step by step. Problem statement, introduction, conclusion and proposed methodology is explained very well. We can also verify results with the help Wordnet. Results are compared very well in the paper.

B. Deep contextualized word representations

In this paper, a deep contextualized word representation model is introduced which is able to model the complex characteristics of word use (syntax and semantics) and how it can be used for vary across linguistic contexts (polysemy). By using this model, they get state-of-the-art results by improving the resulted word representation. The representation of words differ from traditional words embedding in that each representation is assigned to a token which can be used for entire sentence input. Extensive experiments has been done to demonstrate the word embedding representations which works extremely well in practice. They first show that new word embedding can be added easily to existing models for challenging language understanding problems. Strength of the paper is they really addressed the initial problem statement and they developed the model which is able to represent the deep contextualized word representation. Weakness of this paper is that they used many abbreviations but they didn't explain what they are and how they are being used to model the complex use of words. Yes, the methodology proposed in the paper does answer the initial problem statement and they proved the results by performing experiments, training the model on different inputs and testing the model on different dataset. Researchers of the paper tried their best to analyze the results critically but the comparison tables are found ambiguous and it is a tedious task for a reader to understand the comparison table. I found the results of this paper ambiguous and it is very difficult to understand the mathematical equations and tables which are given in the paper and because of this I will say that technical details are correct but it is not making any sense because of ambiguity. I think results from this paper can be verified but it is not an easy task as mathematical equations used in this paper is so complicated. There is no inconsistencies in the paper but ambiguities are found in the paper.

C. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

A new representation model of language called BERT (Bidirectional Encoder Representations and Transformers) is introduced. This paper uses a new model and show some state-of-the art results theoretically in the field if Natural Language Processing like NSP (Next Sentence Prediction), NWP (Next Word Prediction). Weakness of the paper is that the proposed model is yet to be implemented and because of this it is very difficult to verify the data. Experiments performed does answer the initial question that BERT will be able to predict the next word or the whole sentence having state-of-the-art results. Results are critically analyzed with the help of diagrams, graphics. Moreover mathematical equations are also used to prove the results. In the paper, it's also mentioned that with the help of one additional output layer the pre-trained BERT model can create state-of-the-art models for a wide range of tasks. So the appropriate conclusions are drawn from the results. There is a benchmark called GLUE (General Language Understanding Evaluation) and it is a collection of diverse NLP tasks on which different algorithms and models can be tested on. To fine-tune the BERT on the GLUE they used a single sentence as an input and used the final hidden vector according to the first token. If we were to choose from two sentences A and B for a training example, 50% of the time it is a complete random sentence prediction (NSP) and 50% of the time B is the actual sentence followed by A. BERT was also tested on Situation With Adversarial Generations (SWAG) dataset which contains 113K that helps to evaluate grounded inference. Four input sequences are constructed in order to fine-tune BERT on the SWAG dataset. For SWAG, they fine-tuned BERT for 3 epochs having a batch size of 16 and learning rate of 2e-5. Feature-based approach in which certain features are extracted from the trained model has advantage over the classification layer approach .All the technical details are well explained with the help of diagrams and tables. In the paper, results are represented in the tables and graphs and all the results can be verified. Context-sensitive features are extracted from a right-to-left and a left-to-right language model. Concatenation of the left-to-right and right-

to-left representations is the contextual representation of each token.

D. Simple and Effective Multi-Paragraph Reading Comprehension

In this paper, a new method is introduced for paragraph-level question answering models for the case where entire documents are given as input. In this paper, they start by proposing an improved model which will be able to achieve state-of-the-art results. After that to produce accurate paragraph they trained a model which will also be able to achieve high confidence score. They embed words in new vectors by using pre-trained word vectors. After that the next layer receive the concatenation of word-level embedding with character-level embedding. During the training, they are not updating the word embedding. They are using a new method for confidence accuracy which is for multi-paragraph setting by using an un-normalized score given to each step as a measure of the model's confidence. Strength of the paper is that they explained the results very well with the help of graphs and tables. This research paper is fully explained and the results are so clear that's why there is no weakness. Yes, the methodology proposed in the start does answer the initial problem statement. The approach in the paper is critically analyzed with the help of tables. And they performed multiple experiments on each problems with different dataset that's why they get multiple output data and they analyzed the results with others very well .Yes, appropriate conclusions can be drawn from the results. The technical details are correct and they make sense. Yes, the results can be verified by performing the experiments they have performed in the paper. The dataset which is used in each experiment to train the model is given so it is easy to perform the experiment and verify the results. The Paper is very well explained, they have given tables, graphs and diagrams where needed to clear all the ambiguities if the reader has.

E. Overview and Analysis of Existing Decisions of Determining the Meaning of Text Documents

This paper analyzes existing solutions for determining the meaning of text documents. Two most used model and methods are considered which are semantic text processing and the classical process of text in the semantic analyses. In the paper, they found that it is a good approach to develop a single system for solving the problem of analyzing and evaluating texts as well. There are two major directions in determining the meaning of the text and these can be used to solve the problems of the text that is vague and un-meaningful. A lot of work has been done to solve this problem. Strength of the paper is that they explained very well the existing models and methods to understand the meaning of the text documents with the help of experiments on both models with different input dataset and compared the results for accuracy confidence. Weakness of this survey paper is that the benchmark of the input dataset on which models are tested are not provided that's why we can't verify the results. Another weakness of the paper is that the paper was not well written as there were some grammar mistakes. The initial problem statement states that in this survey paper two approaches will be critically analyzed with different experiments and they actually do answer the initial problem statement. Yes, the approach is critically analyzed. Appropriate conclusions are drawn from the results that which model or method for understanding the text is better and why it is better from other. Yes, technical details are correct and they do make sense. Results can be verified if we are given the input dataset they have used in the paper. There is no any serious ambiguity in the paper. All experiments, models, methods and algorithms are explained very well.

F. Information Processing and Retrieval from CSV File by Natural Language

Comma Separated Value (CSV) files are widely used as a fundamental data format. In this paper, a new model is presented which will be used to allow users to easily retrieve information from CSV files using natural language. Deep learning and word representations in vector space are most used techniques in Natural Language Processing. Different modules are developed in the paper for tackling different problems. For conversion processing a new module has been developed specifically for this task and this is very helpful for user as it guide the user to correct the error like unknown words, misspelled words and modify the words sequence. This module can also inform user to remove the ambiguous words that can be segmented more than two different sub-words. Strength of the paper is that they solve the initial problem statement with 97% accuracy. There were only 18 statements where errors occurred. The major cause of the error is the typo error. The weakness of the paper is that they didn't tell the criteria on which input dataset was calculate from different resources. Yes, the proposed methodology actually answer the initial problem statement. The results are critically analyzed for different input datasets with error chances and accuracy percentage. Appropriate conclusions are drawn from the results that on how many sentences their model failed to retrieve the information from CSV file. All the technical details are correct and well explained and it does make sense.

G. Improved Text Language Identification for the South African Languages

To accurately predict the family of language, this paper uses naive Bayes classifier. From this approach, 31% reduction in the language detection error. The mentioned classifier will be useful to create a text language shared identification task for South African languages and will be accurate enough to predict the language family. A classified naive base classifier with various character n-gram text having supervised learning features is introduced which is langid. In this paper, langid is also trained South African Language. In start, the trained classifier has an accuracy of 99.5% as some of the sentences were wrong predicted. But after cleaning the data, the trained classifier had an accuracy of 99.99%. This classifier outperforms the results of other approaches that were developed

before. Strength of the paper is that they get state-of-the-art results using their proposed classifier. Weakness of this paper is that the new proposed classifier is only useful for South African Language. We will need to rewrite the classifier if we want to use it for any other language. Yes, the proposed methodology does answer the initial problem statement. The approach is critically analyzed by giving 2 datasets: first dataset is of correct sentences and second datasets is of wrong datasets so that their classifier can be trained on both datasets and then approach can be critically analyzed. Yes, appropriate conclusions are drawn from the results with all the stats. Results can be verified from the tables and graphs they have provided in the paper. They compared the results with the Google Translate API and their proposed classifier outperforms the Google Translate API results.

H. Text Detection and recognition in Image and video frames

A new method has been represented in this paper for detecting and recognizing text in complex images and video frames. Text understanding from images is performed in a two-step approach where first approach is the detection of the text from image and second is the understanding the meaning of the text that was retrieved from image. The proposed methodology belongs to the top-down category, and consists of two main tasks one is text detection and second is text understanding task. A new multi-hypotheses approach has been proposed to address the text recognition task from images and videos. In this approach, they segmented the image three times and assuming the different number of classes of image each time. Strength of the paper is that they performed the proposed method on multiple images which where noise vary exponentially and their proposed method out performed and get state-of-the-art results. Weakness of the paper is that they used OCR for character recognition but they didn't explained which OCR model they used because there are a number of model that are used for OCR. Yes, the proposed methodology does answer the initial problem statement. The proposed approach is critically analyzed by performing the approach on the different types of images and results are also critically analyzed with the help of tables and graphs. Appropriate conclusion are drawn from the results. Yes, all the technical details given in the paper and it does make sense. All the results can be verified as the proposed method is available on GitHub that's why results can be verified by running the method on different images. There is one ambiguity in the paper and that is the OCR model. They didn't specify which OCR model they used.

I. Translation of natural language queries to structured data sources

Nowadays, most of the people interact with computers and software every day. It is really difficult for a non-technical person to retrieve the data from the structured data source like SQL. In this paper Computational linguistics and natural language processing methods are described. Translation system is developed by Markov Decision Tree. Markov decision can be defined in terms of a state space. The action set that is supposed to be used in the dialogue system includes all possible action it can perform. Prototype of user interface for natural language to database was developed in this research. The structured source data used in this paper is a relational database MySQL which contains the information about the existing solutions, frameworks, models, methods and program libraries that can be used to translate the natural language queries to structured data sources. Strength of the paper is that they really do answer the initial problem statement but there are some constraint that how a natural language sentence should be in order the retrieve the information from structured data source. Weakness of this paper is that they didn't list any result in the table or graph to explain or for comparison. Yes, the proposed methodology does answer the initial problem statement and they implemented the proposed method in C++ which is publically available that's why results can be verified. The approach and result are not critically analyzed as the proposed method was tested on single input rather than multiple inputs. Because of single time execution results cannot be critically analyzed with the result of one dataset. They didn't draw any appropriate conclusion and there was no section of conclusions in this paper. All the technical details correct and they do make sense. As the proposed method is publically available that's results can be verified. There is not any serious ambiguity in the paper.

J. Text Detection and Recognition in Natural Scene Images

This is a challenging task to detect text in natural scenes because of variations in the text size, text-fonts complex background in images. In this paper, a text region detector is designed by using a widely used feature descriptor, histogram of oriented gradients (HOG). Image centroid based distance metric feature extraction system and Zone centroid is used to implement text recognition. Thinning and noise reduction, size normalization, slant correction are done in the pre-processing stage. Thinning and size normalization are very important. As size of the character can vary from font to font that's normalization is very important. A tremendous reduction in text size can be done by thinning. Thinning is used to reduce the thickness of the fonts. To recognize the characters and achieve the state-of-the-art results with 90% accuracy they used a Hybrid type Zone based feature extraction algorithm. Strength of the paper is that they outperform the results with 90% accuracy. Previously the best accuracy in text detection and recognition in natural scene images was 87%. Yes, the proposed methodology does answer the initial problem statement. And the approach is critically analyzed by giving a variety of images as datasets. Appropriate conclusion are drawn from the results and there are other possible interpretations but there is accuracy is less than 90%. The results can be verified with the help of data they have provided. There is a ambiguity in this paper is that what will be the accuracy of the proposed method if someone take random pictures from real time scene.

S No.	Algorithms/Model	Problem	Accuracy
1	Adaboost Classifier Wordnet Principle Component Analysis	Genre classification of books	97%
2	Deep contextualized word representation model	Representation of complex words	Not Given
3	Bidirectional Encoder Representations and Transformers	Unlabeled text to Layers	84%
4	Word-level embedding	High Confidence Score	98%
5	Semantic text processing	Meaning of text documents	98%
6	Semantic Patterns	Retrieve information from CSV	97%
7	Naive Bayes classifier	South African language detection	100%
8	Multi-hypotheses approach with OCR	Text detection from video	Not given
9	Computational linguistics and natural language processing	Natural language to SQL query	92%
10	Hybrid type Zone based feature extraction algorithm	Text extraction from real time scene	90%

Fig. 2. NLP Pipeline.

III. Comparative Study of Related Literature

IV. Conclusions and Future Work

Natural Language Processing is a promising field and different researchers are working on this field from 1962. Natural Language processing will be needed in almost every field. Natural Language Processing can be divided into three steps, one is morphological analysis, second is syntax analysis and third is semantic analysis. Nowadays NLP is considered one of the most difficult problem in AI because nature of the human language makes it difficult. From the literature review it is clear that most of the approaches use deep learning libraries for NLP. Till now naive base classifier is the best classifier for South African language understanding with 97% accuracy and confidence rate. The hot topic in Natural Language Processing is the detection and understanding of text from real-time scene and text extraction from real-time scene is the most difficult filed in NLP. Current work can be extended by thoroughly reading multiple approaches of solving a domain specific problem than those approaches can be compared on the basis of accuracy or dataset. After that all models, algorithms and methods that are used to solve a problem can be compared.

References

[1] Chen, Datong, Jean-Marc Odobez, and Herve Bourlard. 2004. "Text Detection and recognition in Image and video frames." Pattern recognition 37. Elsevier. 595-608.

[2] Clark, Christopher, and Matt Gardner. 2017. "Simple and effective multi-paragraph reading comprehension." arXiv preprint arXiv.

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv.

[4] Duvenhage, Bernardt, Mfundo Ntini, and Phala Ramonyai. 2017. "Improved text language identification for the South African languages." Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). IEEE. 214-218.

[5] Gupta, Shikha, Mohit Agarwal, and Satbir Jain. 2019. "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing." 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE. 269-272.

[6] Peters, MaMark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations." arXiv preprint arXiv.

[7] Pise, Amruta, and S. D. Ruikar. 2014. "Text Detection and Recognition in Natural Scene Images." International Conference on Communication and Signal Processing. IEEE. 1068-1072.

[8] Posevkin, Ruslan, and Igor Bessmertny. 2015. "Translation of natural language queries to structured data sources." 9th International Conference on Application of Information and Communication Technologies (AICT). IEEE. 57-59.

[9] Tapsai, Chalermpol. 2018. "Information Processing and Retrieval from CSV File by Natural Language." IEEE 3rd International Conference on Communication and Information Systems (ICCIS). IEEE. 212-216.

[10] Tereshchenko, Glib, and Iryna Gruzdo. 2018. "Overview and Analysis of Existing Decisions of Determining the Meaning of Text Documents." International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). IEEE. 645-653.