International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Natural Language Processing

Stuti Patel

Department of Computer Science, Parul University

Abstract: ***Natural language processing is widely discussed and researched topic nowadays. As it is one of the oldest areas of research in machine leaning it is used in major fields such as machine translation speech recognition and text processing. Natural language has brought major breakthrough in the field of computation and AI. Various algorithms used for Natural language processing are mainly dependent on the recurrent neutral network. Different text and speech processing algorithm are discussed in this review paper and their working is explained with examples. Results of various algorithms show future scope of research. Natural language processing has not attained perfection till date but continuous improvement done is the field can surely touch perfection line. Different AI now uses natural language processing algorithms to recognize and process the voice command given by user.***

Keywords: NLP, Natural Language Processing, Machine Learning, Machine Translation

1. Introduction

Andrew Ng has long predicted that as speech recognition goes from 95% accurate to 99% accurate. It will become a primary way that we interact with computers. The idea is that this 4% accuracy gap is the difference between annoyingly unreliable and incredibly useful. Thanks to deep learning, We're finally cresting that peak.

Nowadays artificial intelligence is widely discussed buzzwords and is under rapid development. Basically artificial intelligence is a computer program that can do something smart like a human, it is actually machine mimicking human to perform task in his absence and sometimes in better as well as efficient way, broadly speaking

Machine learning is subset of AI. The intelligence of machine is improved using machine learning as through learning algorithm and analysis of different types of data. Deep learning algorithm again and again and improved the machine knowledge according to the output obtained.

Natural language processing is an integral area of computer science in which machine learning and computational linguistics are broadly used. This field is mainly concerned with making the human and computer interaction easy but efficient. Machine learns the syntax and meaning of human language, process it and gives the output to user. The area of NLP involves making computer systems to perform meaningful tasks with the natural and human understandable language.

The reason why natural language processing is so important in future is it helps us to build models and processes which take chunks of information as input and in form of voice or text or both and manipulate them as per the algorithm inside the computer.

Thus the input can be speech, text or image where output of an NLP system can be processed speech as well as written text.

Different algorithms developed to increase the efficiency of processing the language in text form which we are going to discuss here are:

- Long short term memory
- Sequence 2 Sequence model
- Named Entity Recognition model
- User preference graph model
- Word Embedding model
- Feature based sentence extraction using fuzzy interface rules.
- Template based algorithm using automatic text summarization

Similarly, language can be processed even if the input is in speech form. For that various algorithms are developed and the best of them all are:

- Word Recognitions
- Acoustic Modeling
- Connectionist temporal classification
- Phase based machine translation
- Neural machine translation
- Google neural machine translation

In this research paper different algorithm and models are discussed and various improvements done in field of natural language processing. We provide you a basic idea about all the algorithms mentioned above, like on what basis they work on, their efficiency and different applications where these can be implemented for the betterment of the society.

2. Algorithms, models and approach to problem

2.1 Text processing algorithms

1) LSTM:

LSTM stands for long short term memory [1]. Recurrent neural network is chunk of neural network which can remember values.

LSTM are special kind of recurrent neural network which can remember previous input over arbitrary time interval and

predict the output. It is used for training machine through input sets. It is one of the learning models in machine leaning which is broadly used in natural language processing. Stored values are not modified as leaning proceeds. LSTM model is unable to edit the input sets but it can learn from it by computing its frequency according to the event by processing it several times. First step in LSTM is what data input is flushed out of the network. It is decided by forger gate layer '0' represents 'completely forget' and '1' represents 'completely keep'. Next step decides what to store in cell state which is decided by input gate layer. The next layer called a tanh layer creates a vector of new candidate values which is combined to input values to update the state. Some LSTM model have an extra state called peephole connection which lets the gate to check status of cell state and before dropping the data from network.

2) Seq2Seq model:

The tradition seq2seq model contains two recurrent neural network i.e. encoded network and decoder network [2].

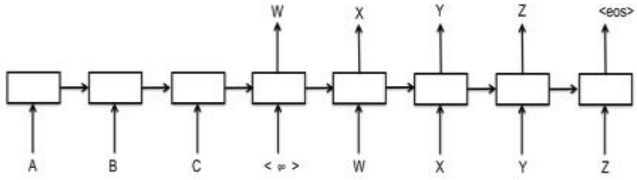
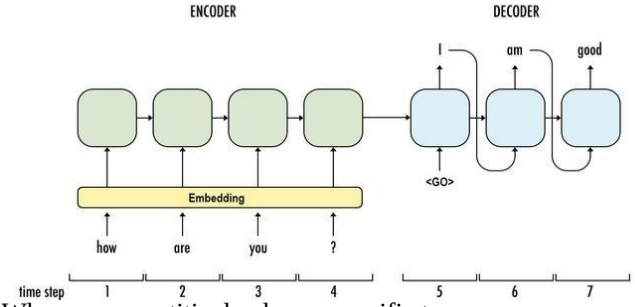


Figure 1: Recurrent neural network structure

Each box represents a RNN most commonly LSTM implemented RNN cell. In this model every input is encoded into fixed size vector which is later processing decoded using decoder.



When user repetitively choose specific tenses

Figure 2: Encoded and decoded structure and working

Firstly, vocabulary list is built and compiled using embedding so that the model can identify the correct grammar syntax. The vocabulary set is processed to check for the occurrences of the words in the vocabulary. The words are them replaced by with ids. Based on id's the reply suggestion is decoded and given as output. Following are some tags used in model compiling the input.

EOS: End of sentence.
PAD: Filler.
GO: Start decoding.
UNK: Unknown: word not in vocabulary.
Following in the examples of Seq2Seq model working:

Question: How are you?
Answer: I am fine.
This pair will be converted to
Question: [PAD, PAD, PAD, PAD, PAD, PAD, "?", "you", "are", "How"]
Answer: [GO, "I", "am", "fine", ", ", EOS, PAD, PAD, PAD, PAD]

3) Named entity recognition Model:

As the name suggests. Name entity recognition is used to identify relevant names classify names by the entity they belong to. NER Model Works in two phases. First speech form [3]. NER model works in two phases. First phase of NER model is to divide the text into segments or chunks to classify them. These chunks are classified in predefined categories such as name of person, organization, location, etc. In form of tokens. The formatting is ignored like Bolding and capitalization. Ex. $ mike^ (ENAMEX, name). In second phase of model. The model can be widely used in language and speech processing, while user preference graph is to be created for smart reply of such suggestions.

4) User preference graph:

User preference graph is used to create a set of user choices. When user repetitively choose specific tenses, Adjectives, Conjunctions and Prepositions etc. A Preference Graph is created so that when user is using similar type of sentences than the model suggests the next word by calculating probability [4].

This words are mapped to each other, hence a preference graph is created for particular user. On big scale implementation of this model, people with similar user preference graph are also grouped together so that the suggestions can have wide scope and variety. This can be implemented in smart reply, smart suggestions, auto reply systems etc.

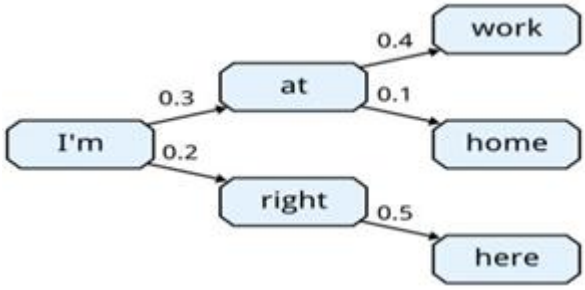


Figure 3: User preference graph example

5) Word Embedding:

Word embedding is derived from feature learning and language modelling in natural process where words and phrases are mapped onto like doors for offer real number of frequency graph.

6) Phrase based machine translation:

Phrase based machine translation is one method of a statistical machine translation. It uses predictive modeling to translate text [5]. These models are created with the help of our learner from a bilingual, large and structured set of texts. With the help of these, the most probable output is created.

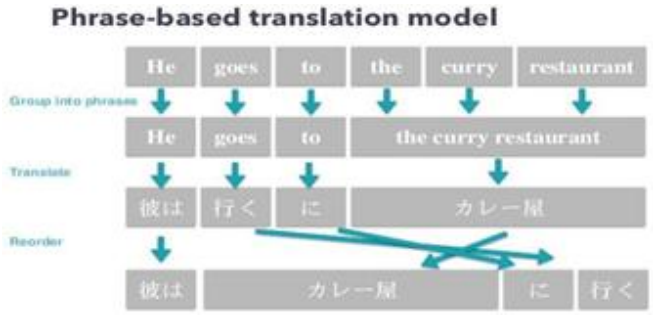


Figure 4: Figure based translation model

The algorithm of PBMT is as follows:

a) Bracket breaking of original sentence into chunks.
b) Find all possible translations for each chunk. In this step, with the help of the corpora, we check how the human translated the text in real world sentence.
c) Generate all possible sentences and find the most likely one. With the help of a different combination of a translation in step B we can generate more than 500 combinations of sentences.
d) Give the probable score by comparing it to training set. Here the training set contains a large database of text from the different books, articles, newspaper etc. By comparing each company combinations of the above. Step to the training data set, we can give it to a probability score. After trying the different combination and passing it through our training data set, we will pick the one that has most likely chunk translation while also having a high likelihood score.

The disadvantage of PBMT [6] is that it difficult to build and maintain. If there is a need to add a new language, then bilingual corpora of that language should be present. For less famous language pair translation trade-offs are made. That is, if translation from Gujarati to. Hindi is does not use a complex pipeline. Instead, it may internally translate it to English and then translate it to Hindi [7].

7) Neural Machine Translation:

Neural transmission translation is newest method to Machine translation. It creates much more accurate translation as compared to the statistical machine translation [5]. Neural machine translation work by sending the input two different layers to be processed before output. NMT is able to use algorithm to learn linguistic rules on its own from statistical models. The NMT system [6] is based on attentional encoder-decoder and operates on sub word units. To improve the efficiency further back translation of the monolingual news corpus is used to as. Additional training data. It is optimal for both direction translations.

The strength of an NMT are that it can better handle verb order forms and avoid verb omissions. It can handle English noun collections. Phrase structure and articles are also well handled by NMT.

[7] The limitation of NMT are ambiguous words into Hindi. It also issues with the forming warp continuous stances dominated dominant problem of MNT prepositions.

2.2 Voice Processing Algorithms

1) Acoustic Modeling:

This contains the references of the individual sounds that make up a word. This individual sound is then assign a label. This label is known as a phoneme. [8] a speech corpus in by using special training algorithm to create statistical AQV presentations which represent each phoneme in Language [9]. This statistical representation is called as Hidden Markov model. Each phoneme has its own HMM. Advantages of using acoustic modeling are that users are motivated to articulate clearly. Smartphones do high quality speech capture; speech transfer to server error free over IP. This model takes lots and lots of training data to create and for many users they work just fine. However, for many other they do not. This is because the data used to generate the model contains sample from tens Of thousands of different speakers, so they are generic. Make specific model for individual is not economical, neither is making models for accents with small population. Acoustic model are limitation of the technology.

2) Connectionist Temporal Classification (CTS):

Connectionist temporal classification is Majorly used for training recurrent neural network. One such example of recurrent neural network is LSTM model. In speech recognition timing is major variable. The input is majorly similar to phoneme, but the timing may be varied. To overcome this problem where LSTM has issues with recognizing phonemes in speech audio. CTC Work by output and scoring, thus being independent of the underlying natural structure. CTC was first introduced in [9].

The CTC network has continuous output which is later fitted through training to model the probability of the label. The output are sequence of labels. If the sequence of label differs only in length then they are considered equal. There are many combination of equal labels, thus making the scoring a non-trivial tasks. [10] paper outlines a dynamic programming algorithm used to compute the sum of probabilities over all path corresponding to a given labeling.

So the algorithm is as follows.

a) Turning sounds into Bits: This is called sampling. By Nyquist theorem, if we sample at least twice as fast as the highest frequency, we can recover the original signal back. For speech recognition, a sampling rate of 16, 000 samples per second is optimal. After this there would be an array of number with each number representing the sound wave's amplitude at 1/16000 a second intervals. We could directly give the sample data to neural network but finding the pattern in such a large dump of data would be difficult and require lots of computations. Increasing the time complexity of the algorithm, so preprocessing is done.
b) Preprocessing our sample data: we reduce the time complexity of algorithm by doing some preprocessing. This preprocessing include grouping of our sample audio into 20 millisecond long chunks. The below image show 1st 20 seconds of audio of first 320 samples.

This short recording is also difficult to process since human speech comprises of a complex sound. There is low pitch sound, mid-range speech sound even some high range

speech sounds. To reduce time complexity further, we use Fourier transform. Thus with the help of Fourier transform, the complex sound wave gets broken into simple sound waves. We then Add up how much energy is present in each one. A spectrogram is created because for the neutral network, finding pattern in the spectrogram is far easier than finding the pattern in a raw sounds file. Below is the representation of above sampled data in spectrogram.

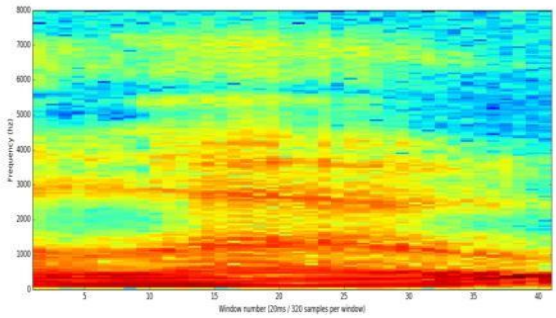


Fig 5. 20ms/320 samples per window

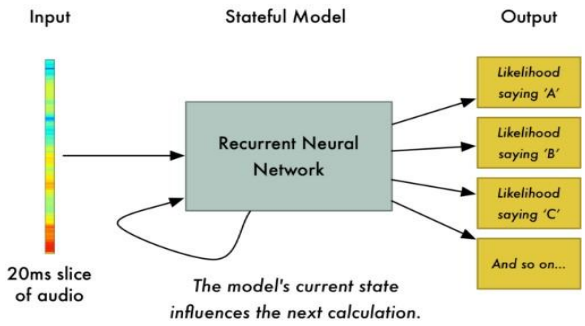


Figure 6: Working of network for speech processing

3) Recognizing character from short sound:

Since the input is straightforward to process. We can feed it to deep neural network. The input being the 20 millisecond audio chunks. For each input, the neural network will try to figure out the phoneme. Send it is recurrent neural network the present output will also influence future predictions. Play. For easy understanding we will reconsider that the input data is of person saying the word "HELLO". [11] so if the prison neural network has recognized "HEL" so far, it's very likely we can say. "LO" rather than some random words such as "XYZ". Since CTC also deals with varying audio length. The output would be as HHHEE_LL_LLLOOO. We will them clean up the output by removing repeated characters. HHHEE_LL_LLLOOO becomes HE_L_LO HHHUU_LL_LLL000 becomes HU_L_LO becomes AAAUU_LL_LLLOO becomes AU_L_LO then we'll remove any blanks: HE_L_LO becomes HELLO. Since all of the sounds similar to HELLO. So these pronunciations based predictions will be continuing with the likelihood score based on the large database of written text. Hence by this likelihood, score of Hello will be greater than the other two. So the output would be shown correctly.

[12] One disadvantage of this algorithm is that if the input audio file is of HULLO. Then the algorithm would not be able to recognize it correctly since the database of written text does not contain more number of achieved HULLO. So the algorithm would malfunction when the readers say word which aren't present in the database of written text.

2.3 Application of NLP:

One of the application of natural language processing which we are going to discuss here is summarization of text automatically with the help of software. We will also discuss two of the best algorithm which were designed to summarize the text and will also compare both of them so as to get the conclusion. But before we discuss about the algorithm, it is better to know more about what automatic text summarization actually is.

Automatic text summarization is basically the task for a software to reduce a large amount of text into meaningful short summary which allow the reader to understand what information the document contains in a short descriptive form as soon as it saves the efforts and times of the user.

There are mainly two general or fundamental way to automatic summarization of text. Those are extraction and abstraction. In text summarization, extractive method, work on choosing between a subset of word, phrase or sentence present in document in its original text to produce an extracted summary. [15] while in. Abstractive method. The algorithm builds an internal semantic representation and then they use natural language general technique. That is, in the technique the machine acts as if it has a human brain and has the ability to generate meaningful summary by understanding the text present in the document. This process creates a summary that is far closer to what a person might actually extract and present as a summary of text. This generated summary includes verbal innovation research by This date have focused primarily on extractive methods which are pertinent for image collection. Summarization, tech summarization, and video summarization.

2.4 Feature based sentence extraction using fuzzy inference rules

The stated algorithm is based on evaluating a sentence in the input data on basis of some rule which categorize the statement and assign those value as low, medium and high. This assignment of the values are done on the basis of rules which are total 8 in count and are known as fuzzy rules or fuzzy logic. These rules are in an if-then form. Like for an example, the algorithm takes a statement F and as an input and apply all the rules on it and assign value to it. Like IF (F1 is H) here it means the importance of the statement on the basis of first logic rule is 'high' similarly all the rules are applied as (F2 is H), (F3 is M), (F4 is H), (F5 is M), (F6 is H), (F7 is M), (F8 is M). [13] and if the statement after being evaluated by the rule satisfies the criteria and is considered as important, then it is added in the summary as per the same sequence as that was in the input data.

The algorithm consists of four stages which process the data and give the final output as a process summary. This stages are first being preprocessing, then feature extraction,

followed by physiologic, scoring and sentence selection and assembly.

Each of these stages consists of sub processes and output of each stage is given as input to the next stage to process.
This algorithm uses certain features to determine the importance of the sentence on which it is included into the summary. The features are:

a) Title feature.
b) Term weight
c) Sentence length.
d) Sentence position.
e) Thematic word.
f) Fuzzy logic

On the basis of these features, the sentence is evaluated by the algorithm and the output is generated.

2.5 Template based algorithm for automatic text summarization

As we saw that in feature based algorithm the sentence are evaluated on the basis of some basic criteria or basic feature and the sentence is arranged in the input data are added as the same in the output summary. While in case of template based extraction algorithm, some modification is being done after extracting the text to arrange the information in the proper and grammar grammatical way to arrange in the summary and to make it more look like a human work. The template based algorithm for automatic text summarization is implemented in two phases.

1) Text preprocessing.
2) Information extraction.

1) Text pre-processing

This part of the implementation includes the module shown below:

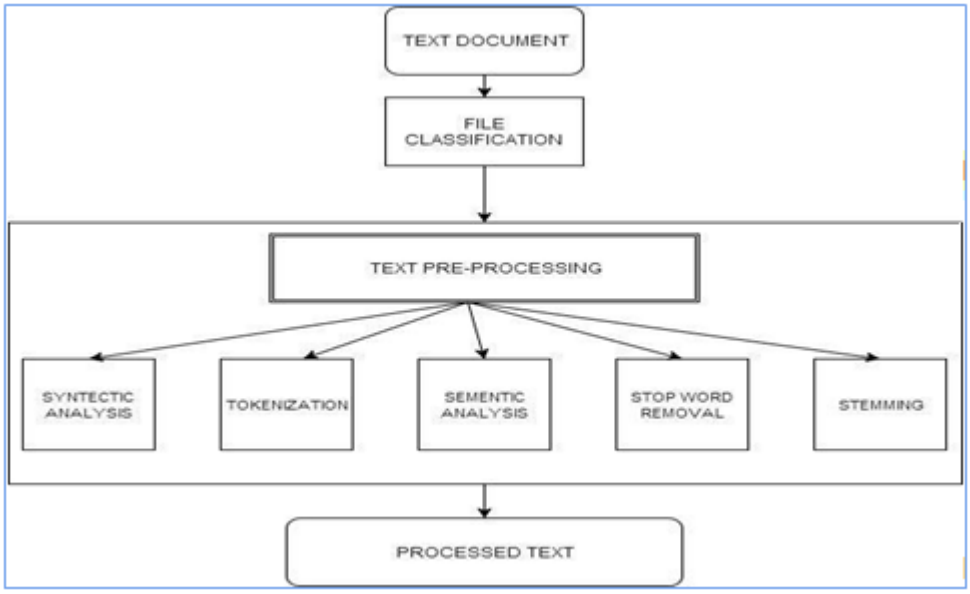


Figure 7: Text Preprocessing model

a) Syntactic analysis

The joy of semantic analysis module is to decide the Starting and ending of the sentence in the input document. Recently, the algorithm takes the full stop symbol as the ending of the sentence and any string to characters up to the full stop symbol is taken as a One full statement.

b) Tokenizer

Job of Tokenizer is to break the sentence given as the output from the semantic analysis module into tokens. The broken pieces of the statement can be words, numbers, or punctuation marks.

c) Semantic

Semantic analysis of phase understands the role of every word which is in sentence. The assignment of a tag is done on every word name as noun, word, adjective and adverb and so on. This process of assigning and dividing the word into different classes is called part of speech tagging.

d) Stop word removal

Some words are used more often in the in the natural language text but their value of importance in extracting meaning is very little in regards when we consider overall meaning of the sentence. Such words are stated as stop words and are removed.

e) Stemming

Steaming is a task of evaluating basic form of Certain word in the input text document. Same word are written in different tenses but in all having the same meaning thus to avoid that, steaming is done and these words having same meaning but different tenses are converted into basic simple tense.

2) Information Extraction

This part of the algorithm includes the following includes:

a) Training the dialogue control
b) Knowledge based discovery
c) Dialogue management
d) Template based summarization

a) Training the dialogue control

During training the system gained knowledge of important or index terms name, body like name of the person, places and temporal formats as per the norms of the rules. Intelligence and efficiency of the algorithm increase with every training set. This is every time we feed the data to the algorithm, it stores the result of the process into the data storage and then uses them as a reference or an experience. But analyzing the in the next input, the concept learned during training are stored in the knowledge base of the system.

b) Knowledge based discovery

Knowledge based discovery here means the process of extracting intelligent information and storing them in an unstructured text form, thus decreasing the need to create multiple storage structures to store different term of different categories, thus reducing the search time and hence improving the overall performance of the algorithm.

c) Dialogue management

The dialogue management module is a process which offers human and computer interaction. With the help of these modules, the user can request the information he needs using normal natural language text. The dialogue control module which is. Linda Bond the training model as the experience data set which except the user request understand it and then refers to its knowledge base and then finally produce the answers which probably contains the sought information.

d) Template based summarization

It is a process of combining together all the meaningful texts present in the input data or the document in a compact format.

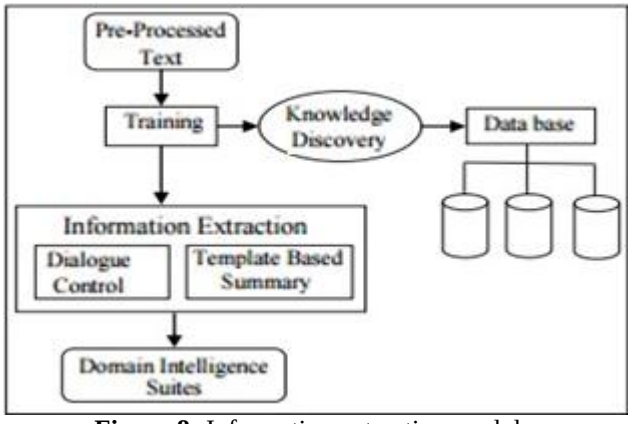


Figure 8: Information extraction module

The algorithm also allow user to prepare the template that has provisions to specify events, locations and name entities, etc. User also has the provision for specifying any number of any type of POS patterns.

2.6 Gap analysis

In template based algorithm, all templates created on different documents by either same user or different user are being stored in the database for future reference as it acts as a training set for the software and hence increasing the accuracy. Of the system, while in few feature based algorithm, there is no concept of database and thus no training set. Hence the gap is being fulfilled. [17]

3. Conclusion

As above context indicate text processing algorithm and based on entity based classification and reference graphs. The text processing algorithm are used in smart reply and smart suggestion in various applications to reduce user's workload and time giving appropriate and efficient output. Whereas in speech processing problem is nowhere must hold but it has improved a lot past decades. Neural and deep learning is used in text processing and speech improve algorithm have resulted in this area. The accuracy level of output is close to perfect due to new improved algorithm which is no close to what human what human would interpret. Various AI are developed based on text processing and speech processing algorithm to assess the user's requirement based on input classification. This improves. Result and user has more personalized result according to his needs. Combination of various text processing algorithm and speech processing algorithms are used for more refined output.

References

[1] Matthew Henderson, Ramial-Rfou, Brian Stropeetal "Efficient Natural Language Response Suggestion for Smart Reply".
[2] Jan Chorowski, NavdeepJaitly "Towards better decoding and language model integration in sequence to sequence models".
[3] Adams Wei Yu, Hongrae Lee, Quoc V. Le "Learning to Skim Text".
[4] FangtaoLi§, YangGao† et al. "Deceptive Answer Prediction with User Preference Graph".
[5] Yonghui Wu, Mike Schuster, Zhifeng Chen et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation".
[6] Luisa Bentivogli, Arianna Bisazza, and Mauro Cettolo "Neural versus Phrase-Based Machine Translation Quality: a Case Study".
[7] Maja Popović "Comparing Language Related Issues for NMT and PBMT between German and English".
[8] Ciprian Chelba "Speech and Natural Language: Where Are We Now and Where Are We Headed?".
[9] Alex Graves, Santiago Fernández et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks" Pittsburgh, Pennsylvania, USA — June 25-29, 2006.
[10] Brian Milch, Alexander Franz "Searching the Web by Voice" Taipei, Taiwan — August 24-September 01, 2002.
[11] Grégoire Mesnil, Yann Dauphin et al. "Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding" IEEE Press Piscataway, NJ, USA Volume 23 Issue 3, March 2015.
[12] Ian McGraw, Rohit Prabhavalkar et al. "Personalized Speech Recognition on Mobile Devices" Shanghai, China 19 May 2016.
[13] MR. S. A. Babar, MS. P. D. Patil "Fuzzy approach for document summarization".

[14] Mr. S. A. Babar, Prof. S. A. Thorat "Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis".
[15] Prashant G Desai, Saroja "A Study of Natural Language Processing Based Algorithms for Text Summarization" Devi Niranjan N Chiplunkar, Mahesh Kini M.
[16] Prashant G. Desai, Sarojadevi, Niranjan N. Chiplunkar "A template based algorithm for automatic text summarization and dialogue management for text documents".
[17] Dipanjan Das Andr´e F. T. Martins "A Survey on Automatic Text Summarization"