Text Normalization and Its Role in Speech Synthesis

Pooja Manisha Rahate, M. B. Chandak

***Abstract:** As the technology is developing day-by-day and most of the human work is done by the machine or systems, it is the need of the today's world to develop systems that can read informal text or words in a proper and standard way even though the format of writing these words or text does not match the standard English words. The informal texts types that exists are the dates, currencies, abbreviations and acronyms of standard words, measurements, URLs, phone numbers etc. This paper focuses on the normalization of such text that converts the informal text into their equivalent standard form which is called text normalization. To produce the equivalent speech form of these non-standard words is the necessity of the today's system. Text normalization is pre-processing step of the natural language processing system. The paper discusses various techniques and methods for the conversion of the non-standard words into standard words. The methods used for classification of the token are regular expressions, used for simple patter match of the token. Naïve Bayes classification for number sense disambiguity and Stochastic Gradient Descent for resolving acronym and class ambiguity .The result and analysis are also mentioned in the form of error-rate of the system, which shows the area for the scope of more improvement in the system.*

***Index Terms:** Naive Bayes, stochastic gradient descent, text normalization, translation memory*

I. INTRODUCTION

Areas such as Information retrieval, Natural language understanding, Text-to-Speech, Automatic Speech Recognition, system for visually impaired persons, Email activators, deals with the real world text in a way or another. All the above systems work on a human or natural language. Natural Language Processing (NLP) is the area that helps to deal with the natural language. It helps the system or machine understand the proper semantics and the syntactic structure of the human readable text and convert or process it in such a way that, the system will be able to read it. Above enlisted areas are the parts of NLP. To process any text into the readable form, it is necessary to first convert the informal text into formal one. In this paper, informal text is addressed using various other names which have the same meaning. It is sometimes addressed as non-standard words (NSW), arbitrary text, un-normalized text, non-English text and so on. The non-standard words are categorized into many classes, such as dates, letter sequences, acronyms, URLs, cardinal, ordinal, telephone number, measurement etc. In this paper total of 16 classes are used for the normalization purpose. These classes are generally called as semiotic classes that work as the identifiers for the NSWs.

Text normalization plays an important role in speech synthesis. Speech synthesis is the representation of text form into the waveform. It consists of two important phases. First, text analysis, where text normalization, phonetic analysis and prosodic analysis is done. And second, waveform analysis which consists of three techniques: articulatory synthesis, formant synthesis and concatenative synthesis. Among these three techniques from waveform analysis, concatenative analysis is commonly used since it uses human pronounced syllable which is stored in the database in the audio form. The corpora used for the project is the Indian-English corpus, there are many challenges that need to be considered for the text normalization. As in India there are many different ways to write a simple NSW. Considered an example of date, which is sometimes written as 1/4/17, 01/04/2017, 1.4.2017, April 4, 2017, etc. Same goes for the money where sometimes the money is written using the (Rs.) symbol or sometimes the devnagri symbol (₹). Therefore, in this work many possible patterns are considered for the identification of the NSW classes. As the data contains Indian-English, there are many words in the corpus that are not present in the standard English dictionary example Indian recipe ingredients names, states, cities and the person names for which the NER (Name Entity Recognition) also does not give the correct result. Thus, there are many problems that need to be resolve for the text normalization of the text. First, there are some words which are the combination of either lowercase letters and uppercase letters or alphabets and numbers. E.g. NC3I (National Command Control Communication Intelligence), GSAT, SAR. The NSWs given above in the example differs in pronunciation. NC3I is the word with the combination of alphabets and numbers where each alphabet needs to be pronounced as a letter sequence with the number between those words. Same rule is applicable for the word SAR (without number). But for the word GSAT, MTOE (million tons of oil equivalent), such words need to identified where the pronunciation changes and we cannot consider the whole word as the proper name (pronounced as a word) and neither we can pronounce the word as letter sequence (G S A T). To pronounce such words the maximum match of the sub word from the last index to the beginning index character must match as the whole word that is to be found in the dictionary. If the sub-word is present in the dictionary then the letters at the beginning of the words must be pronounced as the letter sequence and the remaining word as the whole word. Second, while reading the text, if any acronym occurs in the text,

Manuscript published on 30 July 2019.
***** Correspondence Author (s)

Pooja Manisha Rahate, Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering & Management, Nagpur, India.

Manoj Chandak, HOD of Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering & Management, Nagpur, India.

expansion of the acronym to plays important role for the reader to get the knowledge of the word that what does it exactly means or what is the full form of the read acronym. To resolve this problem, the system will expand the acronym whenever it appears for the first time in the text for the convenient reading. The expansion of acronym will not be done frequently; it will be dependent on its previous appearing index in the text and the current index. Thus acronym expansion will be discussed in detail in the following sections. The output generated by our text normalization system after giving the example paragraph as a input is given below. The system follows the [15] and [2] NSW taxonomy for the normalization of the informal text. The input text given in (1) shows the words in bold that are to be normalized and the output sentence is given in (2) with the bold letters that are normalized.

(1) The actual expenditure of Internal & Extra Budgetary Resources (IEBR) of oil and gas CPSEs in **2016-17** was **₹ 104426.04** crore against Budget Estimate (BE) of **₹ 87214.56** crore. The present capacity of the Refinery is **2.350 MMTPA**. The crude refining capacity utilisation of the refinery was **106.4%** in **2016-17**.

(2) The actual expenditure of Internal & Extra Budgetary Resources (I E B R) of oil and gas **Central public sector enterprises** in **two thousand sixteen seventeen** was **one lakh four thousand four hundred twenty six point four crore rupees** against Budget Estimate (B E) of **eighty seven thousand two hundred fourteen point fifty six crore rupees**. The present capacity of the Refinery is **two point three five zero million metric tonne per annum**. The crude refining capacity utilization of the refinery was **one hundred and six point four percent** in **two thousand sixteen seventeen.**

The remaining part of the paper is organized as follows. Section II discusses the previous work done till now on text normalization and things that are not completely analysed in the other papers. Section III describes the corpora creation, the domain from which the corpus is selected and the NSW classes that will be found in each domain. Section IV describes the NSW taxonomy used in the paper on which the whole classification of the tokens depends. Section V gives the system architecture overview and the components which are present in the system with its small description. Section VI gives the brief description of the methods used while NSW detection, classification and the expansion. Section VII deals with the system performance, which describes the error rate that is obtained in the system predicted classes and expansion with the original token labels and their expansion. Finally, Section VIII ends with the conclusion of the paper and the future scope for the project with a small discussion.

II. LITERATURE REVIEW

Text normalization has been the challenging task in speech synthesis as different languages contains different format or structure of writing text. Thus it is a difficult task to recognize the correct format for normalizing the informal text into formal text. This review may provide some insights to the new researchers and new scholars that will help them to understand different techniques and models that are used up till now for the text normalization research.

Richard Sproat et al. (2001) in his work presented a NSW taxonomy which explains various different categories for the NSW classification. The author used data from new, real estate, recipes and ads domains. The author worked on supervised as well as on unsupervised learning model for the classification for ALPHA class classification. The authors system explains the error rate for the system, which does some mistakes while expanding the NSWs. [15]

Emma Flin et al. (2017) worked on the NSW taxonomy defined by the Sproat et al. (2001). The author made some changes in the taxonomy by adding some more classes to get the more information about the NSW. Author used the Brown Corpus as the dataset for her system. The authors system has four phases that are detection, classification, division and expansion of the NSW. The author used semi-supervised learning for number and alphabets classes. The authors system achieved the accuracy of 91% for the proposed system. [2]

Chen Li, Yang Liu (2014) worked on the Twitter data for the normalization of text. The author used the unsupervised model that creates the lookup table for the NSWs. The model creates a low dimension word embedding for the similarity of the NSWs. The author worked on the sentence level as well as on the work level and achieves different accuracy for the both. The accuracy achieved by the sentence level is better than the word level by 9%. The author used models for re-ranking for achieving the higher accuracy.[8]

Richard Sproat, Navdeep Jaitly (2017) used 3 different models for text normalization; RNN, LSTM and RNN with FST. The author used the dataset constructed by the Wikipedia and hand-labelled the tokens and also write their equivalent output in words. The author worked on English and Russian language and used 16 semiotic classes (NSW types) for the normalization. The model RNN+FST gives the more accuracy in the authors system as FST does not allow the misleading mappings for the words. [5]

Shaurya Rohatgi, Maryam Zare used the dataset of [5] and used the pure deep learning model with the XGBoost classifier for the classification of the NSW token and Sequence to sequence model for the expansion of the NSW. The system defined by the author achieves good accuracy, but as the data gets complex the conversion gets poor for the complex data. The author used vanilla gradient descent without tuning the parameters as an optimiser. [6]

Table I:Non-standard words taxonomy

CLASS	TAG	DESCRIPTION	EXAMPLES
ALPHA	EXPN	Abbreviation	*w.e.f. (with effect from), LKM (line kilometre)*
	ASWD	proper names	*PAN (Permanent Account Number), INDU (Indian National Defence University)*
	LSEQ	letter sequence	*HADR (H A D R), DCN (D C N)*
NUMB	NUM	cardinal number	*89 (eighty nine), 2145 (two thousand one hundred forty five)*
	NORD	ordinal number	*September 21 (twenty first), 7th (seventh),*
	NRANGE	number range	*95-97 (ninety five to ninety seven)*
	NDIG	number as digit	*256 (two five six), 754 (seven five four)*
	NTIME	time	*9:30 am (nine thirty am), 00:15 (twelve fifteen)*
	NDATE	date	*01.2.2017 (first February two thousand seven),25/4 (twenty fifth April)*
	NYEAR	year	*2017 (twenty seventeen), 2006 (two thousand six)*
	NYRANGE	year range	*2017/18, 2017-18 (two thousand seventeen eighteen)*
	MONEY	money	*₹ 416579 (four lakh sixteen thousand five hundred seventy nine)*
	PERCENTAGE	percentage	*25.6% (twenty five point six per cent)*
	NDECIMAL	decimal	*3.6 (three point six), 57.2 (fifty seven point two)*
	NSCI	scientific number	*67.4°N (sixty seven point four degree North), 1Å(one Armstrong)*
SPLIT	SPLT	Mixed	*R-73 (R seventy three), MiG-29 (MiG twenty nine)*
MISC	URL	web address, email	*www.dipp.nic.in,makeinindia-fpi@gov.in*

An improvement to all the work done, we used different methods to identify different NSW categories rather than depending on a single model. For the NSW for which the regular expression gives results for more than one class, to resolve such class ambiguity, we are using Stochastic Gradient Descent with 7-gram words as the feature vector for the algorithm. And for the numbers such as cardinal, ordinal and digit we used Naïve Bayes with collocation vector of +/- three words and their POS tagging.

III. NSW TAXONOMY

After the analysis of data, we found variety of non-standard words categories to which they belong. In this section, we developed a NSW taxonomy which describes the NSWs type and their observed categories. The NSW taxonomy is first developed by Sproat et al. (2001) [15] for the corpora. The same taxonomy is used by Emma Flin et al. in [2] where the author modified the existing taxonomy according to the corpora used in the project. The taxonomy described in this paper is the combination of both the taxonomies and one or two more classes were added according to the data found in the corpora. The brief description of the NSW taxonomy is described below in the following Table 1.

There are four main classes in the NSW taxonomy; ALPHA NUMB, SPLIT and MISC. The ALPHA class consists of three tags; ASWD, EXPN, LSEQ. Any NSW that contains only alphabets belongs to ALPHA class. The ASWD tag specifies the word needs to be pronounced as a proper name even though the word is the acronym or initialisms, because the word itself is present in the dictionary and in the pronouncing dictionary. The tag LSEQ specifies the token to be read as a letter sequence. The token with the EXPN tag specifies that the token needs to be expanded. The NUMB class contains several tags to identify the number token. For the tag NUM, NORD and NDIG we used the Naïve Bayes classifier as the pronunciation of these tags depends on the contextual information. The new tag is added in the taxonomy NYRANGE, specifies the year range, as the year range needs to be pronounced in a different way where we do not replace the symbol – or / with the word 'to'. The SPLIT tag is used to notify the system that, the token need to be split for the pronunciation. MISC class contains the URL tag which is used for labelling the web URLs and email addresses.

IV. CORPORA

A. Domain Description

The corpus is created on the Indian-English text which contains many non-standard words which are not present in the standard dictionary and their pronunciation is also not present in the CMU Pronouncing dictionary. The corpus contains the text from annual reports of Defence Ministry, Petroleum Ministry, the data from Indian news headlines and the data of recipes. The reason behind using the text data from different domains is that, it contains different formats to non-standard words and contains single acronyms with two or more meanings or expansions.

Defence Ministry: The text written in the annual reports of defence ministry is in the systematic and well-edited format. This data is used to identify the acronyms in the corpus as it contains number of various acronyms and date formats.

Petroleum Ministry: The reason for selecting petroleum data is, first since it is the government report the text is well organized and well formatted. Second, it contains acronyms that are present in defence but are having different meaning in this data related to petroleum. Thus the system will try to identify the correct expansion for the acronym according to the context where it is used.

It contains many currency values and different date formats than the defence data.

Recipes: The recipes data is used for identifying the fractions from the text and the measurements. The data is useful to identify the ambiguity for the number that are written using the oblique symbol. Such numbers create ambiguity such that whether to identify the focused token as a date, simple fraction value or as a year.

AstroSat: The astrosat data is the book data which describes the astronomy mission of the ISRO organization. The purpose behind using this data is, it contains many measurement values and Standard units that must be correctly identifiable by the system.

Indian News: The use of Indian news data is, it contains dates with the fractional format and the timing of each news headlines. This date and time formats create ambiguity with the fraction number and time creates ambiguity with the ratio value such that, the focus token or word must be pronounced as ratio or as a time.

Each corpus plays their individual roles for token identification according to the context word the NSW contains. Total of 2 lakh tokens are present in the database for the training and testing purpose. The no. of NSWs in each dataset with respect to the total tokens of individual dataset is shown in Table 2.

Table 2: Size of each corpora and number of non-standard words detected

Corpus	Defence	Petroleum	Recipes	AstroSat	News
Total # tokens	35K	64K	53K	15K	33K
# NSWs detected	2K	7.5K	4K	2.5K	3K

For evaluation purpose the 20% data from each domain is used which will not be used for the training purpose. The test data will be new for the system, to check the system accuracy and overall performance on the unseen data. The distribution of token classes in the particular domains is given in the following table with the count of each class for their respective domains.

B. *Preprocessing of Corpora*

In this step, the hand labeling of each token of the corpora is done. The hand labelled corpora is represented as the original or true corpora which contains correctly labelled tokens. This work is done manually; as such there is no other way to identify the correct label of the tokens without providing the base knowledge or prior knowledge for the token classification.

The labeling of the token is done according to the taxonomy described in the above section. For the ambiguous tokens, the surrounding words are considered for applying the label for the focused token. As the corpora contain data from different tokens, the processing of each domain corpus is done individually. But for testing the data from all domains will be combined for measuring the system accuracy and performance. All the data is stored in the single file and each domain data is labelled with its appropriate domain name. The purpose of storing the data in a single file is that, the corpus contains acronyms whose expansion depends on the domain data. The details of which will be explained briefly in the further sections.

Thus, the original corpora will be used for the evaluation purpose of the predicted data. Using the original tagged tokens and the token label generated by the system, the error rate of the system will be calculated. This error rate will explain the performance of the system and the misclassification done by the system for labeling.

V. SYSTEM DESCRIPTION

In this section, the components of the system are described in detail and the role of each component is explained.

A. *Architecture Overview*

The system designed for the text normalization contains many sub-components for successful normalization of the NSW. There exists three main components in the system, that are; NSW detection, NSW classification and NSW expansion. The architecture of the system is diagrammed in Fig. 1 and includes the following components;

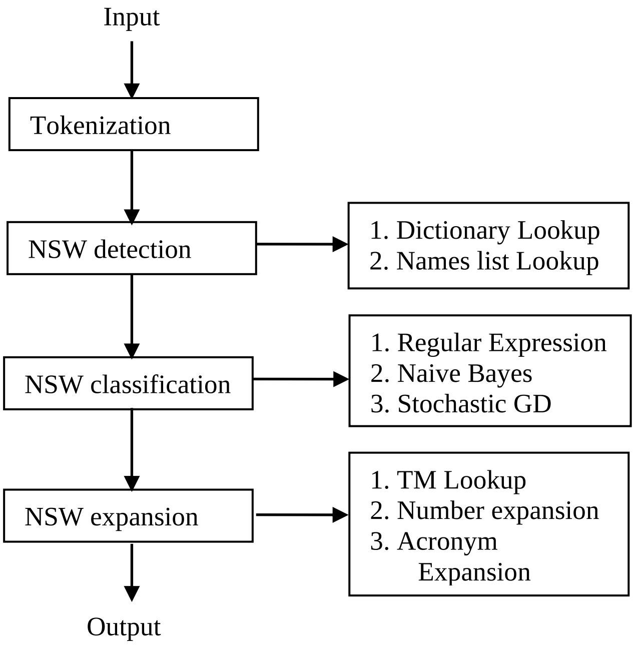

Fig. 1: Architecture of the text normalization system.

Tokenization: In tokenization phase, the text breaks down using the whitespaces and the beginning and trailing punctuations are removed.

NSW detection: In detection phase, each token is cross checked with the words present in the dictionary. If the word is present, the token is kept as it is; otherwise the token is processed to the next phase for classification.

NSW classification: The output from the detection phase is the input of the classification phase. In this phase the NSW token is processed and the tag is applied to the NSW according to the NSW taxonomy.

NSW expansion: In this phase the NSW is expanded according to the tag specified in the previous phase. The expansion is done by using parallel corpus for ALPHA class and for NUMB class the classes are defined for expansion.

VI. METHODOLOGY

A. *Text Preprocessing*

The pre-processing is first phase of text normalization system. In this phase two methods are performed; tokenization and splitting. As we are using the Indian-English data, there are many words and the acronyms that are not present in the CMU dictionary and in any standard dictionary. Therefore the dictionary creation is done from the vocabulary of the dataset.

1. *Tokenization and NSW detection*

For NSW detection, two dictionaries are maintained; one dictionary that contains the list of English words and Indian words; and second contains the dictionary of names of the person and location. In tokenization method, all the text data is first separated using the whitespaces. After the generation of tokens, the data is passed to the splitter. The splitter is the module that separates the beginning and trailing punctuations from the tokens. After splitting the punctuations, each token from the beginning is checked against the word in the standard dictionary and the names dictionary. If the token is present in any of the dictionary given above, then the token is left as it is and no operations are performed on it. If the token is not present in the dictionary lookup, then the token is labelled as the NSW and the classification techniques are performed on it. The classification techniques are described in brief in the next section.

B. *NSW classification*

The classification module works on the three techniques regular expressions, Naïve Bayes (NB) and Stochastic Gradient Descent (SGD). All the techniques are briefly described below.

1. *Regular expression*

The regular expressions (RE) are used to find the pattern of NSW classes. The classes such as dates, acronyms with full stops, URLs, email addresses, numbers with per cent symbol, numbers with decimals, money symbols, fractional numbers will be identified using the regular expressions. Total of 12 regular expression patterns are used to identify the NSW pattern. The RE are designed as functions in the system.

2. *Naïve Bayes for number sense ambiguities*

The Naïve Bayes classifier is used to identify the ambiguous senses for the number classes i.e. cardinal, ordinal and digit class. Consider an example number 25; here according to the taxonomy the number 25 can be read as a twenty five if its class is cardinal, twenty fifth if the class is ordinal and two five if the class is observed to a digit class. All these classes identification depends on the words surrounded by the number token. Therefore instead of writing the hand written rules for class identification for number, the Naïve Bayes classifier is used.

For the NB classifier, the feature vector is created to calculate the probability of for each sense (denoted by s) for a given token. The probability equation for Naïve Bayes is as follows;

$$\hat{s} = argmax_{S \in senses} P(s) \prod_{i=1}^{n} P\left(V_w^i \middle| s\right) \qquad (3)$$

where V_w is the feature vector,
s represents the sense of the number (cardinal, ordinal and digit)

The feature vector used in the above equations is created using the collocation vector with +/- 3 words and their POS (part of speech) tagging. While creating the feature vector the punctuations are not included if they appears as the tokens instead of that punctuation token the word previous or next to it is used. The feature vector is created using the training data which contains columns as the collocation vectors and rows contain the number sense. The row of the whole feature vector matrix represents the sense for the observed collocation vector. The senses with the maximum probability will the class for the focused number token. In this way, the class for the ambiguous number class is classified.

3. *Stochastic Gradient Descent for class ambiguities*

There are some NSWs whose structure and representation of writing is same, but there pronunciation changes according to the words that are surrounded to them. Consider some cases, such as NSWs 9:30, 2017-18, 2/5, 2:3, ½, 95-97, 2017/18 and so on. All the NSWs that are considered above, has different pronunciation according to the context words among them. The NSW 9:30 could be considered as a time or it can be a ratio of any mixture also. 2017-18 represents range most of the time but 95-97 could be represented as a money range or year range depending on its context. Same goes for the 2017/18, 2/5, ½ where according to the corpora we are using 2017/18 is not the fraction but represents the financial year, where 2/5 could be represented as the date or could be a fraction number depending on the context. To resolve this class ambiguity, stochastic gradient descent classifier gives correct output for the focused token.

Stochastic gradient descent is the optimization algorithm which uses linear regression for classification purpose. In [6], the author used the ASCII value to represent each character in the token whereas in [16], the author uses the Unicode value to represent each character of the token. In both the paper, the authors used the XGBoost classifier and worked at the character level for creating features for the algorithm. But in this paper, we used a unique number value to represent each word in the corpora and for every NSW that will be replaced by some special tag to be identified easily. Thus to number each word in the corpora, we have created a combined list of all the unique words from all the four domains and give number to each words. In SGD algorithm, instead of training the whole dataset for only few classes, we used +/- 7 n-gram with relevant its label (i.e. the class name of NSW for the given context words). The dataset for SGD algorithm contains 800-1000 rows where all the possible context words combination for the particular class is considered. The unique numbers then get replaced by their respective words in the data. Thus the range of numbers extends to 7-8 digits long according to the length of the words, we performed feature scaling on the dataset used for the SDG, so that all the numbers are set into the specific range.

The results obtained using the SDG are as expected. It gives the accuracy of 95% on the test data using the context words and the label as the input to the classifier.

C. NSW expansion

1. Translation Memory Lookup

Translation memory (TM) is the memory that is used as the parallel corpus for the acronym, abbreviation and symbol. The TM is the XML specification file which contains several tags to store the information about the data. The translation memory used in the project contains the acronyms and abbreviations found in the dataset with its extension. The TM lookup is used when the token is identified as any currency or measuring symbol or units, abbreviation or any acronym. The reason for storing the acronym full-form in the TM will be described in the next subsection. Each NSW contains the label tag which contains the NSW class name. Suppose, their occur any NSW token which is identified as the abbreviation, then in the TM, only those tags will be searched for the expansion whose label contains the abbreviation. This will save time for searching and retrieving the data from the TM.

2. Number Expansion

Several functions are defined for the number expansion. According to the class identified for the number, the function will be called from the set of functions which is defined as *Fn*, such that, $Fn = \{f1, f2, \ldots, fn\}$. For the fractional number, the function $f \in Fn$ will be called where the symbol (/) will be replace by the word (by). E.g. token 5/6, identified as fraction, pronounced as *five by six*. For the numbers such as year 2006 and 1965, the pronunciation of both the numbers is different. Thus we have defined some rules for such year numbers;

- if the first and last index of the year contains a number from the range [1-9] and the middle two index contains zero (0); then the number will be pronounced as number value thousand number value.
- if there is zero present on the 2nd index then partition the number in pair of 2 for pronunciation.
- if there is zero present on the 3rd index from the beginning of the year, then considered first 2 digits as a single number and the last number as a single number.
- if no zero is present in the year then also divide the number in the pair of two for pronunciation.

In this way, the expansion of the number is done differently for different NSW classes.

3. Acronym Expansion

In the work done up till now on text normalization, the authors pronounced an acronym as sequence of letter every time it appears in the data. In this project, we will expand an acronym also so that while reading any acronym in the text the full form of the acronym must be known to the user which gives reader the knowledge and more information about the topic to the user. As per the human tendency, while reading any text if any acronym is identified while reading, we try to find out its full form such that, what exactly does it mean. Same approach is used in this project. This is the reason why we are storing the acronyms full forms in the Translation memory.

To expand the acronym several things are taken into consideration. Consider two example sentences;

i. John lives in USA.

ii. The President of USA is going to visit Indian Prime Minster.

In the first sentence, while reading USA we read it as letter sequence U S A. But while reading the second sentence we pronounce the full form of USA i.e. United States of America. If we observed both the sentences, we will find that the USA in first sentence lies in the Verb Phrase (VP) of the sentence, while the word USA lies in the Noun Phrase (NP) of the sentence. Thus this is our baseline for the expansion of the acronym in the text. The sentence will be chunked in NP and VP parts and then the expansion of acronym will be decided. But this goes for the single sentence, what if we have a whole paragraph or page with some information. The acronym used in that text if the text is about some company or some organization, the same acronym may appear again and again in the Noun Phrase, and it will not be relevant to expand the acronym every time. This will create disturbance will listening to the text. Thus to overcome this problem, we set the value for the index range for the acronym occurrence. If the acronym is written in the round brackets followed by its fullform, at that time we will read that acronym as sequence of letter. But if the acronym occurs without followed by the expansion and is in the NP chunk, we will expand it. If the same acronym appears within the range of 50-75 words, then the acronym will not be expanded and it will be read as sequence of letter, otherwise the acronym will again be expanded. The question that again arises here is, how do we calculate the difference between the index of the previous acronym and the current acronym. For this purpose also we used the translation memory. Along with the acronym and its expansion, we stored the index value of the acronym that appears every time in the system. When the acronym appears more than one time, the previous index value is used from the TM for the acronym to calculate the difference between the current and previous acronym index. If the difference is below 75 then acronym will not get expanded, otherwise the acronym expansion is done. In this way the system normalized the informal text after the processing performed in the system.

VII. SYSTEM PERFORMANCE

This section discusses the overall system performance for the designed system. For the evaluation of the system, the tested data is compared with the original data which contains the hand labelled tagging and the expansion of the tokens.

Here we have evaluated the system performance differently for the classification of NSW and for the expansion of NSW. The accuracy achieved by the classification phase in 94%. As in the classification models we used +/- 3 words for the number classification, if the window size gets increased, the classifier can collect more information for predicting the tag for the number class. In stochastic gradient descent for the tag classification for ambiguous classes we use the +/- 7 words for the tag prediction. The SGD in the project uses the default parameters for the classification. But if we tune the parameters and more examples were added in the feature vector, the SGD classifier will outperform the results. For the expansion of the NSW, the accuracy of 98% is achieved as the expansion is done using the dictionary lookup for ALPHA class, and for NUMB class the functions are written. The accuracy for the MISC class for URL token is somewhat less because; in the URLs and email addresses the words are written in the compact way which makes the system difficult to find the maximum match for the words such that how much characters to be taken to form a word from the dictionary and which characters must be pronounced as a letter sequence.

VIII. CONCLUSION AND FUTURE SCOPE

The paper presents the work done for the text normalization and the role it plays in the speech synthesis. Text normalization is the important part for any Natural Language Processing application because; if the normalization of the word is not done correctly and properly, it may convey the wrong message to the listeners which will provide the bad impact on the application by the users. The paper uses the Sproat et al. (2001) and Emma Flin et al. (2017) NSW taxonomy for the NSW categories found in the corpus. The corpus is created using the five domains that are Indian Defence and Petroleum Ministry annual reports, ISRO AstroSat corpus, Indian Recipes corpus and Indian News. In this project the work up till the expansion of the NSW is done. The output normalize text is not converted into the speech waveform because; as the data used is the Indian-English, the CMU Pronouncing dictionary does contains the pronunciation for the Indian-English words.

For the future scope of the project we can create a pronouncing database for the Indian-English words that contains all the pronunciation for Indian organization names, person names, food etc. This database after creation can get integrated with the system to generate the speech waveform form for the input text tokens.

As the text normalization is the trending topic in the NLP areas, many other techniques can be used for the normalization of the text where the no rule-based system will be used and the corpus is given directly to the model for the conversion of informal text to formal text. But to create such model human efforts are again needed to create a hand labelled corpus that will be given as the input to the system and the system itself will learn the classification rules for the tokens as many different languages exists in the word and to create a labelled corpus for each and every language will be a time consuming and need experts to create such corpus.

REFERENCES

1. Emma Flin et al, "A Text Normalization System for Non-Standard EnglishWords",Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 107–115, Copenhagen, Denmark, September 7, 2017.
2. Meenakshi Sharma,"Text Normalization Using Hybrid Approach", International Journal of Computer Science and Mobile Computing, Vol.4 Issue.1, January- 2015, pg. 544-554.
3. Hay Mar Htun, Theingi Zin, Hla Myo Tun, "Text To Speech Conversion Using Different Speech Synthesis", NTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 07, JULY 2015.
4. Anand Arokia Raj et al, "Text Processing for Text-to-Speech Systems in Indian Languages", 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.
5. Richard Sproat, Navdeep Jaitly, "RNN Approaches to Text Normalization: A Challange", arXiv preprint arXiv:1611.00068, Jan 2017.
6. Shaurya Rohatgi, Maryam Zare, "DeepNorm - A Deep Learning approach to Text Normalization", arXiv preprint arXiv:1712.06994, Dec 2017.
7. Chen Li, Yang Liu, "Improving Text Normalization via Unsupervised Model and Discriminative Reranking", Proceedings of the ACL 2014 Student Research Workshop, pages 86–93, 2014.
8. Dileep Kini, Sumit Gulwani, "FlashNormalize: Programming by Examples for Text Normalization", Proceeding IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence, Pages 776-783, 2015.
9. Suhas R. Mache et al, "Review on Text-To-Speech Synthesizer", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 8, pg. 54-59, August 2015.
10. Lopez Ludeña, V., San Segundo, R., Montero, J. M., Barra Chicote, R., & Lorenzo, J. (2012). "Architecture for text normalization using statistical machine translation techniques." In IberSPEECH 2012 (pp. 112 – 122). Madrid, Spain, 2012.
11. Conghui Zhu et al, "A Unified Tagging Approach to Text Normalization", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pg. 688–695, Prague, Czech Republic, 2007.
12. Chris Lin, Qian (Sarah) Mu, Yi Shao, "iTalk A 3-Component System for Text-to-Speech Synthesis", unpublished.
13. Slobodan Beliga, Miran Pobar and Sanda Martincic-Ipšic, "Normalization of Non-Standard Words in Croatian Texts", arXiv preprint arXiv:1503.08167v2, 30 Mar 2015.
14. Gokul P., Neethu Thomas, Crisil Thomas and Dr. Deepa P. Gopinath, "Text Normalization and Unit Selection for a Memory Based Non Uniform Unit Selection TTS in Malayalam", Proceedings of the 12th International Conference on Natural Language Processing, pg. 172-177, Trivandrum, India, 2015.
15. Richard Sproat et al, "Normalization of non-standard words", Computer Speech and Language (2001) 15, pg. 287–333.
16. Subhojeet Pramanika, Aman Hussaina, “Text Normalization using Memory Augmented Neural Networks”, arXiv preprint arXiv: 1806.00044v2, July 2018.
17. Daniel Jurafsky, James Martin, Speech and Language Processing, Pearson India, 2nd Edition.

AUTHORS PROFILE

Pooja Manisha Rahate received Bachelor's from Jhulelal Institute of Technology, Nagpur, India in 2017 and pursuing Master's degree in Computer Science and Engineering from Shri Ramdeobaba College of Engineering, Nagpur, India. Published a case study on Comparative Study of String Matching Algorithms for DNA dataset in IJCSE Journal in May 2018. Participated in IEEE Region10 student branch website contest and is a IEEE member.

Dr. Manoj B Chandak is Professor & Head of Department, Department of Computer Science & Engineering in Shri Ramdeobaba College of Engineering and Management, Nagpur, India. He holds the Ph.D. and having 25.1 years of teaching experience. He is awarded as the 2nd Merit for B.E. Computer Technology and as a 1st Merit for M.E. Computer Science and Engineering. He is the member of IEEE and IEEE Computer Society.