Implementation of Context Window and Context Identification Array for Identification and Interpretation of Non Standard Word in Bengali News Corpus

Chandan Kundu, Rajib Kumar Das, Kalyan Sengupta
EIILM
India
chandankundu2008, rkdas70@gmail.com, kalyansen@iiswbm.edu

ABSTRACT: *Non Standard Word (NSW) identification and its interpretation is a major challenge in information retrieval system. Real text not only contains ordinary words and names but also contains non-standard "words" including numbers, abbreviations, dates, months, currency, amounts and different types of numbers. In reality, we can not find NSWs in a dictionary, nor can one find their pronunciation by an application of ordinary "letter-to-sound" rules. Non standard words have greater inclination towards ambiguity in terms of their interpretation and pronunciation than ordinary words. So it is very much required to identify and interpret the NSW properly while we are going for further analysis of text. In real text, NSW are represented in diversified formats. It could be represented by digits, words or combination of both. There are different ways to identify and interpret the NSW such as n-gram language model, decision tree, supervised and unsupervised techniques, weighted finite-state transducers and pattern identification using regular expression.*

In this paper we presented a novel work to identify and interpret the non standard words in Bengali news corpus. We introduced the coupling of the normal text normalization using the three components of analysis, viz. (i) generation of optimized regular expression for different semiotic classes, (ii) consideration of optimized context window size, and (iii) employ the concept of context identification array (CIA).

Keywords: Non Standard Word (NSW), Context Identification Array (CIA), Context Window (CW), Regular Expression (RE)

Received: 1 October 2013, Revised 12 November 2013, Accepted 17 November 2013

© 2013 DLINE. All rights reserved

1. Introduction

Information retrieval is the process of extraction of information from structured and unstructured data. In this process, the information that is needed to be extracted is presented in the form of a query and this query tries to match against the information contained in the database. Much of the research and development in information retrieval is aimed at improving retrieval efficiency [6] [7].

Different fields of language and speech technology must deal with real text. In some cases the dependency is indirect: automatic speech recognizers normally depend on language models that are trained on text. In other cases the dependency is direct: for instance, machine translation, topic detection or text-to-speech systems start with text as their input.

Unfortunately, written language deviates from this ideal in two important ways. First, in most of the languages there is ambiguity even for ordinary words: if we write *book*, it is up to you as the reader to figure out from the context whether we meant *book* as a manuscript, or *book* to reserve (ticket or hotel reservation). Secondly, many things in the text are not ordinary words. These

include: numbers and digit sequences of various kinds; acronyms and letter sequences in all capitals; mixed words (win98); mixed case words (*WinXP*); abbreviations; Roman numerals; universal resource locators (URLs), e-mail addresses and so on.

Non standard word identification and interpretation [8] [9] is a type of information retrieval technique that is widely used in natural language processing. Normally dictionary contains meaning or interpretation of standard words namely nouns, verbs, adjectives etc. In contrary, non standard words are not standard words and their meanings or interpretations are absent in the dictionary. But there are occasions where meaning or interpretation of NSW is important. This paper addresses the design and implementation of NSW identification and interpretation, formally known as text normalization system [5]. The experiment has been carried on a set of Bengali (Bangla) news corpus.

Proper deciphering of the NSWs in a text can be extremely useful to many NLP applications such as Text to Speech (TTS) system. Text to Speech (TTS) system [4] is one of the major areas of research in natural language processing. At the beginning of the TTS processing, we are required to submit raw texts including NSW as input and based on given input the TTS will convert it into spoken word. So it is important to find out the NSWs and they must be converted into corresponding meaningful words according to their context.

2. Related works

Considerable works have been done by different researchers on different languages likely English, Chinese, Japanese, Hindi and other languages. But insignificant amount of work has been done on internationally popular language like Bengali (*Bangla*). The basic idea behind the NSWs identification and interpretation are more all less same involving tokenization, token classification, token sense disambiguation and standard word generation to get normalized text. Individual step differs in way of implementation. Different researchers, however, adopted different ways for tokenization (whitespace-separated tokens [5] [11], end of sentence [4] [11], use of Flex [1], Regex [11], break iterator [11], etc), token classification (ngram language models [5], decision trees [5], regular expression [4], Bayesian classification [12], etc) and token disambiguation (*n*-gram sequence model [14], statistical techniques [13], decision-list [9], etc).

[1] Reported the accuracies for *Cardinal-Number* and *Year* were 96.1%, and for *Float* and *Time* were 98.6%. [2] found overall accuracy for Japanese and Chinese were 72.9%. From the study of [3], we can point out the following accuracy levels - *Floating point*-100%; *Currency* -100%; and *Time* – 62%.

Comparatively low accuracy levels of certain semiotic classes demand further investigation and analysis. It may be pointed out that only a few researchers used the concept of *context* in their analysis [10]. But along with this, concept of *Context Identification Array* (*CIA*) as proposed by us may be a unique idea to tackle the issues.

In natural language processing, context of the word plays a vital role while interpreting the meaning of the sentence. Considering the context of the NSWs, they can be categorized into different semiotic classes namely *Time, Currency, Abbreviation, Date & Month, Telephone number, Year, Number, E-mail, URL* etc. While converting these texts into speech, we have to concentrate more on interpretation rather than the orthographic symbols, so that in Bengali ''১২.৫ মি.'' (12.5mi.) (for full details of Bengali words vide Appendix A) would sound as 12.5 মিনিট (12.5minute) but ''মি. রয়'' (mi. Roy) would sound as "*Mister Roy*".

The following examples may establish the fact that interpretation of a NSW is subject to depend on context in the text.

1) The NSW '1998' could be considered as a *YEAR* (pronounced as nineteen ninety eight) or a *NUMBER* (pronounced as one thousand nine hundred ninety eight) or telephone extension number (pronounced as one nine nine eight).

2) মি. (mi.) could be considered as MINUTE or MISTER (as discussed above).

3) In Bengali, "কাজটা ২.৩০ থেকে ৪.৩০ মিনিটের মধ্যে শেষ করবেন" (finish the work anytime between 2.30 and 4.30 minute), first set of digits 2.30 could be treated as either *FLOATING NUMBER* or *TIME*, giving rise to ambiguity.

Although the second set of digits is directly associated with *time*, which is absent in the first set.

4) In Bengali, 'টা' / 'টে' / 'টো' (ta/te/to) could be interpreted frequently as *TIME*, for example, "বিকেল 5 টা" (afternoon 5), but it may sometimes indicate *QUANTITY* as well, for example, "5টা বই" (5 books).

From the above examples it is observed that for proper pronunciation or interpretation, in Bengali, context of the word is very important.

In this paper we have concentrated on exploring the context of the NSW, so that we can identify and interpret the NSWs more accurately compared to other ordinary NSW identification techniques.

3. Methodology of extraction and suitable identification of NSW

Text normalization is a multi steps process. In this paper we have divided the whole processes into different sub processes namely (i) construction of Regular Expression (RE) (ii) Classification of similar types of NSWs, namely "7টা" (7 ta) considering *quantity* or "7টা" (7 ta) considering *time*, based on REs (iii) construction of *Context Identification Array* (CIA), (iv) identification of the optimum *Context Window* size for successful performance (using training data), (v) final (or confirmed) classification for proper identification of a specific class of NSW and (vi) converting the NSWs of numerical forms to textual forms.

It is worthy to mention that individual programming module (date & month, email, money etc.), regular expression and CIA are used for different semiotic classes. It is important to note that single programming module is used for semiotic classes having same type of nature (e.g., time-quantity-float, year-number-telephone number etc.) but different regular expressions and CIAs, suitable for those semiotic classes, are used. Structure of individual module varies as per structure of different semiotic classes. The whole process is depicted in the following figure:

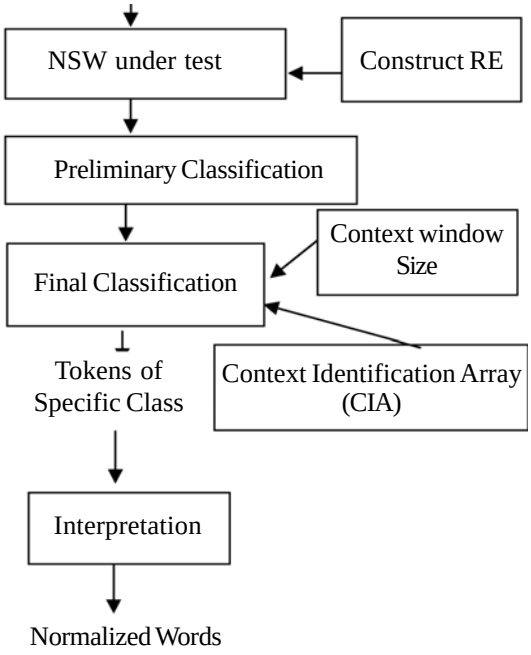

Figure 1. Flow chart for Bengali Text Normalization

3.1 Difficulties of unified model

Bengali is an inflected language. Bengali literature follows proper rules while constructing sentences and punctuations. But in most news corpus, sentences and punctuations are used with little variations. While we try to find out any word, generally we focus on prefix or suffix and this fact is applicable for any language. In case of *time* and *date*, to mention the range, we use "২.৩০ থেকে ৫.৩০মিনিটের" (from 2.30 to 5.30 minute) and "12 আর 14ই জানুয়ারি" (12 and 14th January). Here we notice that for both the examples, in first set of digits, we have not used any suffix (i.e., minute or month etc). We can mention more examples where either prefix or suffix is absent. Sometimes we come across the situations when prefix or suffix is absent as the writer mentioned them in the previous sentences, e.g. "ভারত ১৯৮৩ সালে প্রথম ক্রিকেটে বিশ্বকাপ জেতে।তারপর আবার জেতে ২০১১ তে।" (India first won the world cup cricket in the year 1983. Again they won the cup in 2011).

The following table (Table 1) shows the frequencies of normal syntax and deviated syntax (as described above), which is known

as inflected feature of the language, as extracted from a relatively voluminous training data set[1] (2.03 MB) of a Bengali news paper, *Anandabazar Patrika*[2]. The regular expressions and Context Identification Arrays (CIA) are developed manually by analyzing this data set prepared from news corpora.

Category	Normal (%)	Inflected (%)
Time	98.8	1.2
Year	96.2	3.8
Float	100	0
URL	100	0
Date	99.4	0.6
Money	100	0
Telephone no.	98.1	1.9

Table 1. NSW variations in Bengali News corpus

For some semiotic classes, final classification (Figure 1) is not mandatory because presence of certain symbols directly identify the semiotic classes as in the case of email that is identified by '@' symbol or in the case of percentage that is recognized by the symbol '%' and so on.

Because of the above facts, we have decided to use the algorithm with little variations for different semiotic classes irrespective of the use of different regular expression.

3.2 Algorithm

The algorithm 2 depicts the working principle of our proposed technique.

The algorithm presented here is a straightforward rulebased approach that uses exhaustive search procedures both in preliminary and final classification stages. In the preliminary classification stage, we have used regular expression for individual semiotic class. CIA plays a vital role in the final classification stage. After getting the patterns from the preliminary stage, left and the right windows have been taken care of considering the digit or word format of the digit of that pattern, for example, '7টা' (7ta) must be broken into '7' and 'টা'(ta). Now we start comparing every token of the left and right windows, starting from the digit ('7'), with the tokens of the CIA. If there is a match between any token of the respective windows and CIA, we can conclude the semiotic class of that NSW.

Nevertheless, the algorithm needs slight changes in case of NSWs having identical representation, for example, timequantity, year-number-telephone number etc. In these cases, search operation is made with each CIA one after another, for example, tokens of left and right context windows of digit '7' are compared first with time CIA and then with quantity CIA. If there is a match between the contexts of the digit and more than one CIA tokens, then we calculate the distances from the digit to the contexts present in the context windows. The digit, which is close to the context, is identified as NSW of that semiotic class. But in case of tie (distances of same length), we can not conclude properly the semiotic class of that digit and accuracy levels go down.

3.3 Construction of RE and its importance in NSW identification

A regular expression (RE) is a set of characters that specify a pattern. Regular expressions are used when we want to search for specific lines of text containing a particular pattern. We can think of regular expressions as wildcards on steroids. In NSWs identification technique, RE plays a vital role considering NSW as specific pattern.

In this paper, we have presented REs that have designed after a thorough study of Bengali news corpus. The study reveals that

[1]Available at http://eiilm.co.in/Training_Data_Set.txt

[2]Available at http://www.anandabazar.com/

Step 1: Open the file in read mode and read the file line by line.

Step 2: Check for specific pattern/s using regular expression (re) for specific semiotic class/s in each line.

Step 3: Split the line.

Step 4: Find number of resulted pattern/s extracted using re in each line.

Step 5: Select number of iteration required for same type of structure of different semiotic classes (time-quantity, yearnumber-

telephone number).

Step 6: Iterate for individual resulted pattern of each line for final classification.

Step 7: Split the pattern and get the index of the digit or word format of the digit.

Step 8: Set the flag to zero.

Step 9: Iteration based on left Context Window size.

Step10: Check start of the line. If it is, then got to previous line and continue the iteration according to the length of the left context window.

Step 11: Iteration based on the length of the Context Identification Array.

Step 12: If there is a match between any element of CIA and tokens within the context window size then stop the iteration. Make the flag to one. Based on the value of step 5, either go to step 6(for more than one similar structure) or stop the iteration (only one structure) because we have got the NSW and go to step 18.

Step13: If there is no match then start iteration based on Right Context Window size.

Step14: Check end of the line. If it is, then go to next line and continue the iteration according to the length of the right context window.

Step15: Iteration based on the length of the Context Identification Array.

Step16: If there is a match between any element of CIA and tokens within the context window size then stop the iteration. Based on the value of step 5, either go to step 6(for more than one similar structure) or stop the iteration (only one structure) because we have got the NSW and go to step 18.

Step 17: If there is a match between the contexts of the digit and more than one CIA tokens, then we calculate the distances from the digit to the contexts present in the context windows. The digit, which is close to the context, is identified as NSW of that semiotic class.

Step18: Convert the digit into its corresponding word format

Step19: End of algorithm

Figure 2. Algorithm for finding NSWs using Python style (only main part is given)

some accessory texts and punctuation marks are associated with the desired pattern and without these accessory texts and punctuation marks we can not predict the NSW accurately. For example, "২.৩০ থেকে ৪.৩০ মিনিটের" (from 2.30 to 4.30 minute) gives us two NSWs namely 2.30 and 4.30. But if we consider only digits for the identification of '*time*' NSWs, we will get NSW '4.30' only, because it is only preceded by the context 'minute'. The digit '2.30' will not be considered as *time* NSW as it is not preceded by any *time* context, but in reality it is *time* NSW. So, here, we have to consider the text 'থেকে' (fromto). In another case, "২-৩ টে বই" (2-3te boi i.e. 2-3nos. of books) should give us two NSWs namely 2 and 3 if and only if we consider the punctuation '-'. Therefore proper construction of REs helps us to search particular pattern more accurately. Keeping in view of the above discussion, we have designed REs for different semiotic classes; some of which are listed in the Appendix C.

3.4 Preliminary Classification

In the preliminary classification phase, we are searching for specific pattern using constructed REs. In this approach, we are not using the *tokenization* process in a strict sense as an initial phase. Instead, we extracted the desired tokens of different semiotic classes using well organized REs .Thus we have tried to eliminate the time that is required for *tokenization* phase. We have used Python functions for searching specific pattern line by line. After this phase, tokens of same type are being classified. In this

classified group, tokens of same type but of different semiotic classes are being kept. For example, 'টা' /'টে' / 'টো' (ta/te/to) are kept in a single class but they may be categorized either into *time* or *quantity or float value.*

So, we can say the preliminary classification phase is a screening phase before the final classification.

3.5 Final Classification through Context Identification Array (CIA)

Here, we have performed the final classification using the concept of '*Context*' of a word. In any language, context of the word has a significant role to understand the perspective or background of that word. Context of the word precisely identifies the actual meaning of the word because we have found strong coupling between the context of the word and the meaning of that word. For example, in English, if you write '23rd', it does not imply anything alone. 23rd may signify quantity or month or something else. But if there is word '*January*' after or before 23rd, then we can precisely say that 23rd means 23rd January. In another example, 1972 alone does not mean anything; it could be year or ordinary number or anything else. The word '*year*' along with 1972 specifically helps to understand we are talking about '*year* 1972'. Likewise, there are ample of examples which sufficiently prove the need of '*context*' of a NSW.

In this paper, to achieve a higher accuracy, we have considered two new concepts implemented jointly, namely, *Context Identification Array* (*CIA*) and *Context Window* (*CW*).

3.5.1 Construction of Context Identification Array

Thorough study of news corpus reveals that there are some words that present before or after the NSWs and presence of those words identify the nature of the NSWs. CIA is a special kind of array which stores words or patterns i.e. '*contexts*' that are used to identify the NSWs precisely: $CIA = \{P_0, P_1, P_2, \ldots, P_n\}$ (where n represents the size of Context Identification Array). Therefore construction of CIA is very important because ultimately it segregates the NSWs of same nomenclature into different classes. A list of CIA for Bengali language is illustrated in Appendix B.

It is required to mention that, in case of YEAR, RE and CIA are almost similar because we have chosen RE in such a way that minimizes the search area.

3.5.2 Importance of Context Window

As we have mentioned earlier, final classification is carried out by means of *Context Identification Array* and *Context Window* (*CW*) [1]. In this experiment CIA and CW are complement to each other because both of them are used simultaneously to justify the context as well as the nature of the NSW. CW identifies the region within which there is a possibility to find out the context of a NSW. Each token within the CW undergoes checking to see whether it corresponds to specific '*context*' or not. CW comprises of windows namely *left Context Window:* $l\_c\_w = \{t_0, t_{-1}, t_{-2}, \ldots\ldots, t_{-n}\}$ and *right Context Window:* $r\_c\_w = \{t_0, t_1, t_2, \ldots, t_n\}$ (where *n* represents the size of context window; array starts at the index value of the target token). If the context presents in the left window, then it is known as *Left biased semiotic class* (e.g. *time*, *telephone number* etc) and if the context presents in the right window, then it is known as *Right biased semiotic class*(e.g. *Date & Month, Quantity* etc. Total window size is equal to sum of left and right window sizes.

Size of CW varies for different semiotic classes. Intense study of the news corpus reveals that context of a NSW must present within a specific region centering the NSW. Let us explain the situation with a suitable example. In case of *Date & Month*, name of the month normally appears within the range of 1-5 tokens i.e. "জুন মাসের ৫ তারিখ (CW size 2), জুন মাসের ১ থেকে ৫ তারিখ" (CW size 2 and 4 respectively), ৫ই জুন (CW size 1) and so on. But in a true sense CW size is not rigid. Selection of CW size depends on the knowledge of the researcher on the literature. If the CW size is not selected properly, there may arise several problems, such as

• Arbitrarily long CW may select number that is not belonging to that semiotic class.

• Small CW size may omit the context and the number without this context may be categorized to faulty semiotic class.

Generally, long CW size requires more iterations, however if the context is arrested early, the iteration will be stopped immediately.

3.5.3 Implementation of CIA and CW

In the final classification stage, we have considered the output of the preliminary stage, where each classified group comprises

of same type of tokens but of different semiotic classes. At first, each individual token of a single line is taken under consideration and the index value of each one is identified in the given line. After getting the index value, we have considered the left context window l_c_w = $\{t_0, t_{-1}, t_{-2}, \ldots\ldots, t_{-n}\}$ and right context window r_c_w = $\{t_0, t_1, t_2, \ldots, t_n\}$ starting from that index value. CIA comprises of m number of context patterns, $CIA = \{P_0, P_1, P_2, \ldots\ldots, P_m\}$. Now we start comparing each element of the left CW and also the right CW with the individual token of the CIA. If there is a match between the token of the CIA and the token either to the left or to the right CW, then we can understand the nature of the target token to be classified to a specific semiotic class:

$$flag = \begin{cases} 1, \text{ if CIA } [P_i] = = \text{l\_c\_w } [tj] \text{ where } i = 0, 1, 2, .., n \text{ and} \\ \qquad\qquad j = 0, -1, -2, \ldots, -n \\ 1, \text{ if CIA } [P_i] = = \text{l\_c\_w } [tj] \text{ where } j = 0, 1, 2, .., n \text{ and} \\ 0, \text{ otherwise} \end{cases}$$

Total tokens extracted using RE = 390					
Actual number of Date & Month NSWs present in the data set = 99					
Serial No.	**Left Window**	**Right Window**	**Total length**	**Time (sec)**	**No. of actual tokens**
1	0	0	0	4.19e- 06	0
2	0	1	1	4.19e-06	0
3	0	2	2	2.512	92
4	0	3	3	2.59	92
5	0	4	4	2.72	95
6	0	5	5	2.78	95
7	0	6	6	2.90	97
8	1	2	3	2.64	92
9	1	3	4	2.68	92
10	1	4	5	2.79	95
11	2	3	5	2.78	92
12	2	4	6	2.93	95
13	2	5	7	2.98	95

Table 2. Experimental result for Date & Month with variable length of Context Window

where flag = = 1 indicates the receiving of the confirmed result.

The token then undergoes for conversion from digit to word representation based on its semiotic class.

4. Results

The experiment has been carried out on a test data set[3] (803 KB) which is different from the training data set, prepared from Bengali news paper *Anandabazar Patrika*. The following tables show the experimental results executed on two semiotic classes namely *Date & Month* and *Time* respectively.

In case of *Date & Month*, database contains exactly 99 NSWs belonging to this semiotic class. Using regular expression, in the

[3]Available at http://eiilm.co.in/Test_Data_Set.txt

preliminary classification stage, we got 390 tokens. As all of them do not belong to the *Date & Month* semiotic class, we performed final classification stage. From the experiment, it has been observed that for some context window sizes (serial nos. 1-4, 7-9, and 11 in Table 2), we got results that are not optimum. Serial nos. 5- 6, 10, 12, 13 (Table 2) deliver us much better results which is 95. It is interesting to mention that in serial no.7, we got 97 tokens but two of them do not belong to the *Date & Month* semiotic class. From the Table 2, we can identify that the optimum window size to be 0-4 (left context window size 0 and right context widow size 4), as shown in Table 4.

Total tokens extracted using RE = 485					
Actual number of Time NSWs present in the data set = 101					
Sl. No.	**Left Window**	**Right Window**	**Total Length**	**Time**	**No. of actual tokens**
1	0	0	0	3.08	0
2	1	1	2	3.12	37
3	1	2	3	3.13	56
4	2	2	4	3.22	59
5	2	3	5	3.23	72
6	3	3	6	3.41	82
7	3	4	7	3.49	91
8	4	4	8	3.5	91
9	5	4	9	3.51	95
10	5	5	10	3.54	95
11	5	6	11	3.57	95
12	6	5	11	3.58	95
13	6	6	12	3.59	95
14	6	7	13	3.63	95
15	7	7	14	3.64	95
16	8	8	16	3.65	104

Table 3. Experimental result for Time with variable length of Context Window

Semiotic Class	**Left Context Window Size**	**Right Context Window Size**
Date & Month	0	4
Time	5	4
Year	2	4

Table 4. Optimum Context Window Size

On the other hand, database contains precisely 101 NSWs belonging to *Time* semiotic class. The preliminary classification stage gives us 485 tokens. Using final classification, we could identify maximum of 95 out of 101 NSWs that has been shown in the serial nos. 9-15 of Table 3. Other than these serial nos., we would not get optimum number of *TIME* NSWs. It has been observed from the Table 3 that the optimum context window size is 5-4 (left context window size 5 and right context widow size 4) as shown in Table 4.

The study reveals that the search operations of NSWs with respect to different semiotic classes, some semiotic classes achieve exactly 100 % accuracy whereas some fails to get that. The following table carries the comparative details of findings between the early system [3] and our proposed system [Table 5].

The accuracy level for *Money, Telephone No., URL, Percentage and Quantity* is 100%. It has been observed in case of *Date & Month, Year* and *Time*, accuracy values are close to 100% but not 100%. This shortfall is due to the inflective nature of Bengali language. In case of *Time*, we overcome the problem that has been faced in [3], for example, " কাজটা 10.15 থেকে 11.15

Semiotic Class	Accuracy of the early system[4] in %	Accuracy of the proposed system in %
Date & Month	71.02	95.95
Money	100	100
Telephone No.	100	100
Year	82.45	97.54
Time	63.57	94.05
URL	100	100
Percentage	100	100
Quantity	72.61	95.34
Float	100	100

Table 5. Accuracy values for different semiotic classes

মিনিটের মধ্যে শেষ করেবন" (Finish the work between 10.15 and 11.15 minute). Here first float number 10.15 and second float number 11.15 have been identified as *time* because the context '*মিনিটের*' (*minute*) is present within the three and one context window sizes respectively and hence we solve the problem of [3]. But in some cases we could not identify the *Time* NSWs properly, for example, "আমি বিকেল ৪টে বাড়ি ফিরব। তারপর ৫টা নাগাদ খেলতে যাব।" (I will come back to my house at 4 in the evening. Then around 5, I will be out to play). In this case '4' is identified as *Time* because it is associated with '*evening*' context in the first line. But in the second line, with continuation of the '*evening*' context, '5' is not associated explicitly with any *Time* context. So we could not identify the '5' as *Time*, instead it has been identified as Quantity, as 'টা' ('*ta*') comes after '5'. We have faced similar type of problems in case of Date & Month and Year also. Hence the accuracy levels are slightly less than 100%. By incorporating more complex rules and techniques, we can overcome this problem.

Thus, after getting different types of NSWs, we have interpreted according to their natures. In case of Year, 1972 will be interpreted as "উনিশশো বাহাত্তর" (nineteen seventy two). But if it is Money, then it will be "এক হাজার নয়শো বাহাত্তর" (One thousand nine hundred seventy two). In case of Telephone No., it will be "এক নয় সাত দুই" (one nine seven two).

5. Conclusion and Future Work

In this paper we developed a model for identification of NSWs in Bengali news corpus. It is interesting to observe that Context Identification Array (CIA) and Context Window (CW) size are important determinants for identification. However, size of the CW depends on the type of the semiotic classes and the complexity of the language. This complexity carries a direct relationship with inflected nature of the language, Bengali language to be particular. In case of Bengali news corpus, the degree of inflation is quite high and therefore we could not 100% accuracy for all semiotic classes like, '*year*', '*time*', etc.

The research opens up a number of facets of further research by different approaches and methodologies e.g., maximum entropy, Bayes theorem, SVM etc. Once efficiencies of these methods are assessed, the optimal technique can be addressed for some particular applications on Bengali news corpus.

References

[1] Panchapagesan, K., Talukdar, P.P., Sridhar Krishna, N., Bali, K., Ramakrishnan, A. G (2004). Hindi Text Normalization. Fifth International Conference on Knowledge Based Computer Systems (KBCS), Hyderabad, India.

[2] Olinsky, C., Black, A.W (2009). Non-Standard Word and Homograph Resolution for Asian Language Text Analysis**.** ICSLP2000, Beijing, China, p. 733- 736.

[4] The accuracy figures for early system correspond to our test data along with our constructed REs. However, there has been a little improvement achieved in accuracies because of detail designed and well organized REs.

[3] Alam, F., Habib, S. M. M., Khan, M. (2009). Text Normalization System for Bangla. Proc. Conference on Language and Technology (CLT09), NUCES, Lahore, Pakistan, January, p. 22-24.

[4] Xydas, G., Karberis, G., Kouroupetroglou, G. (2004). Text Normalization for pronunciation of Non- Standard Words in an Inflected Language. *In*: Proceedings of the 3rd Hellenic conference on Artificial Intelligence (SETN04), Samos, Greece.

[5] Sproat, R., Black, A.W., Chen, S., Kumar, S., Osetendorfk, M., Richards, R. (2001). Normalization of non-standard words. Computer Speech and Language, p. 287–333.

[6] Göker, A., Davies, J. (2009). Information Retrieval: Searching In the 21st Century. John Wiley & Sons Ltd.

[7] Feldman, R., Sanger, J. (2007). The Text Mining Handbook. Cambridge University Press.

[8] Cavnar, W. B., Trenkle, J. M. (1994). N-Gram- Based Text Categorization. *In*: Proceedings of SDAIR- 94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, UNLV Publications/Reprographics, p. 161–175.

[9]Yarowsky, D. (1996). Homograph Disambiguation in Text-to-Speech Synthesis. Progress in Speech Synthesis, Springer-Verlag, p. 158-172.

[10] Raj A. A., Sarkar T., Pammi S. C., Yuvaraj S., Bansal M., Prahallad K., Black, A.W. (2008). Text Processing for Text-to-Speech Systems in Indian Languages. ISCA SSW6, Bonn, Germany, p. 188- 193.

[11] Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S. (2000). A unified POS tagging architectureand its application to Greek. *In*: proceedings of the 2nd LREC, Athens, Greece, p. 1455-1462.

[12] Golding, A. (1994). A Bayesian hybrid method for context-sensitive spelling correction. *In*: proceedings of the 3rd workshop on very large corpora, p. 39-53.

[13] Luk, A. K. (1995). Statistical sense disambiguation with relatively small corpora using dictionary definitions. *In*: proceedings of the 33rd annual meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, p. 181-188.

[14] Jelinek, F. (1997). Statistical Methods for Speech Recognition. MIT Press, Cambridge.

Appendix A

Bengali words, pronunciations and meanings

Bengali word	Pronunciation	Meaning
মি.	Me.	Mister (to designate a male person) / short form of minute (to represent minute)
মিনিট	Minute	Minute
রয়	Roy	Surname
কাজটা	Kajta	The work
২.৩০	2.30	2.30
থেকে/ হ ইতে	Theke/ hoite	From
৪.৩০	4.30	4.30
মিনিটের	Miniter	Of minute
শেষ	Sesh	End
মধ্যে	Modhya	In between
করবেন	Korben	You Will do
টা / টে / টো	Ta/te/tow	Suffix used to repres ent time and quantity
আর	Arr	And
১৪ই	14e	14th

ভারত ১৯৮৩ সালে প্রথম ক্রিকেটে বিশ্বকাপ জেতে।তারপর আবার জেতে ২০১১ তে।	Bharat, unissho tirashi, sale, prothom, crickete, bissyocup, jete, dari, tarpor, abar, jete, duhazar, agaro te, dari	India, 1983, in the year, first time, in cricket, world cup, won, end of sentence punctuation, then, again, won, in 2011, end of sentence punctuation
২-৩	Dui-tin	2-3
বই	Boi	Books
জুন মাসের ৫ তারিখ	June, maser, panch, tarikh	June, in the month, 5th, date
আমি বিকেল ৪টে বাড়ি ফিরব। তারপর ৫টা নাগাদ খেলতে যাব।	Ami, bikel, charte, bari phirbo, dari, tarpor, panch ta, nagad, khelte, jabo	Me, in the evening, at 4 pm, house, return, end of sentence, then, at 5 pm, around, to play, will go
উনিশশো বাহাত্তর	Unissho bahattor	1972
এক হাজার নয়শো বাহাত্তর	Ak hazar noysho bahattor	1972
এক দুই সাত দুই	Ak di sat dui	1972
এবং / ও	Abong /O	And
শে ,ই ,লা,রা,ঠা	Sh-a, e, law-a, rawa, tha-a	Suffix used to represent date
সন, সালে	Son, sale	Year, year
শতাংশ	shatangsho	percentag
টা. , টাকা	Ta., taka	Indian currency in short form, Indian currency
পয়সা , প.	Poisa, po.	Indian currency, Indian currency in short form
অ-য়	Oa - untostoao	Bengali alphabets
'জানুয়ারি','ফেব্রুয়ারি','মার্চ', 'এপ্রিল','মে','জুন', 'জুলাই','অগস্ট', 'সেপ্টেম্বর','অক্টোবর','নভেম্বর','ডিসেম্বর')	January, February, March, April, may, June, July, August, September, October, November, December	January, February, March, April, may, June, July, August, September, October, November, December
সকাল/সকালে ,বিকালে/বিকাল, সন্ধ্যা/ সন্ধ্যে, দুপুরে/ দুপুর , রাতে/ রাত/রাত্রে , ভোরে/ভোর	Sakal/sakale,Bikal/ bikale, Bela, sandhya/ sandhye, dupure/ dupur, rate/rat/ratre, bhore/bhor	Morning, after noon, noon, evening, after noon, night, dawn
বত্সর,বছর,বত্সরে,বছরে,সাল,সালে, শতক,শতকে ,সন, সনে	Batsor,bachar,batsore,sal,sale,shatok,shatoke,son,sone	year

Appendix B

Structure of Context Identification Array to find the different NSWs in Bengali news corpora

Semiotic Class	CIA
Date & Month	if context is in ('জানুয়ারি','ফেব্রুয়ারি','মার্চ', 'এপ্রিল','মে','জুন', 'জুলাই','অগস্ট', 'সেপ্টেম্বর','অক্টোবর','নভেম্বর','ডিসেম্বর') then NSW implies Date & Month else NSW implies any Number
Time	if context is in ('সকাল' , 'সকালে', 'বিকালে' , 'বিকাল', 'বেলা' ,'সন্ধ্যা','সন্ধ্যে', 'দুপুরে' , 'দুপুর' , ' রাতে' , 'রাত', 'রাতে', 'রাত্রে' , 'ভোরে','ভোর') then NSW implies Time else NSW implies Quantity
Year	if context is in ('বত্সর','বছর','বত্সরে','বছরে','সাল','সালে', 'শতক','শতকে' ,'সন','সনে') then NSW implies Year else NSW implies any Number.

Appendix C

Structure of Regular Expressions for different semiotic classes

Class	RE
Date & Month	(r'\s+\d{1,2}[-/.]\d{1,2}\s+ | \s+\d{1,2}\d*\s+[হ]*[ই]*[ত]*[ে]*[থ]*[ে]*[ক]*[ে]*[এ]*[ব]*[ং]*[ও]*\s\d+\.*\d*[শ]*[ে]*ই*[ল]*[া]*[র]*[া]*[ঠ]*[া]*| \s+\\d+\.*\d*\s* [শে]*ই*[লা]*[রা]*[ঠা]*\s+^\d{1,2}\,?\s+|\s+\d{1,2}\,?\s+| \(\d+\.*\d*\s')
Year	(r'সন\s\d+ | [\']\d*\sসাল[ে]* | সাল[ে]*\s\d* | \d*\sসন[ে]* | \d*\sসাল[ে]*')
Time	(r'\d+\.\d\d\s[ম][ি]\.*[ন]*[ি]*[ট]*|\d+\.*\d+\s[ট][া][র]| \d*\.*\d*\d*\s[হ]*[ই]*[ত]*[ে]*[থ]*[ে]*[ক]*[ে]*\s\d*\.*\d* \d*\s[ট][া][র]*\s| \d+\s[ট][া] \s|\d*[\,]*\d*[\.]\d+')

URL	(r'http://[a-z]+[0-9]*[.][a-z]+ [0-9]*[.]*[a-z]*[0-9]* |www[.][a-z]+[0-9]*[.][a-z]+[0-9]*[.]*[a-z]*[0-9]*')
email	(r'[a-z]+[0-9]*[_.-]*[a-z]*[0-9]*[a-z]*@[a-z]+[0-9]*[.] [a-z]*[a-z]*')
Percentage	(r'[0-9]+[.]*[0-9]*\s*% | \d+[.]*\d*\s*শতাংশ')
Phone	(r'\+*\d*\d*\s*\(*\d*\) *\-*\s*\d\d \d\d\d\d\d\d|\(+\d{4}\-\d{4}\)+')
Money	(r'\d+[\,]*\d*[\.]*\d*\s*[\ট][\া] [\.]\s\d*[\,]* [\.]*\d*\s \প.| \d*[\,]* [\.]*\d*\s\টাকা\s\d* [\,]*[\.]*\d* \s\পয়সা| \d+[\,]* [\.]*\d*\sটাকা|\d+[\,]*[\.]* \d*\s\পয়সা|\d+\s[প][.]| \d+[\,]*\d*[\.]*\d*\s[\ট][\া][\.]| \Rs.\s\d+[\,]*\d*[\.]*\d*[\/-]* |[অ-য়]+\s*[অ-য়]+\s*টাকা| [অ-য়]*\s*[অ-য়]+\s*টা[.]+ |[অ-য়]+\s*পয়সা|[অ-য়]+\s*প[.]+| [অ-য়]+\s*[অ-য়]+\s*টাকা \s*[অ-য়]+\s*পয়সা|[অ-য়]+ \s*টা[.]+\s*[অ-য়]+\s*প[.]+| \s\d+\.*\d*\s+[ক]*[ো]*[ট]*[ি]*\s* [ল]*[ক্ষ]*[্য]*\s*[ল]*[া]*[খ]*\s*[হ]*[া]*[জ]*[া]*[র]*\s* [শ]*[ো]*\s*[ক]*[ো]*[ট]*[ি]*\s*টাকা| [অ-য়’ ঁ]+ \s+[ক]*[ো]*[ট]*[ি]*\s*[ল]*[ক্ষ]*[্য]*\s* [ল]*[া]*[খ]*\s* [হ]*[া]*[জ]*[া]*[র]*\s*[শ]*[ো]*\s*[ক]*[ো]*[ট]*[ি]*\s* টাকা')