ISSN 2277-3061

Text Normalization for Telugu Text-to-Speech Synthesis

Dr.K.V.N.Sunitha, P.Sunitha Devi
Principal BVRIT Hyderabad, College of Engg. for Women,
Bachupally, Hyderabad, A.P. India.
k.v.n.sunitha@gmail.com
Assistant Professor, G.Narayanamma Institute of Technology & Science
for Women, Shaikpet, Hyderabad, A.P. India
sunithareddy.katta@gmail.com

ABSTRACT

Most areas related to language and speech technology, directly or indirectly, require handling of unrestricted text, and Text-to-speech systems directly need to work on real text. To build a natural sounding speech synthesis system, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text. A novel approach is used, where the input text is tokenized, and classification is done based on token type. The token sense disambiguation is achieved by the semantic nature of the language and then the expansion rules are applied to get the normalized text. However, for Telugu language not much work is done on text normalization. In this paper we discuss our efforts for designing a rule based system to achieve text normalization in the context of building Telugu text-to-speech system.

Keywords

Speech Synthesis, Classification, Token Sense Disambiguation, Text Normalization.

Academic Discipline And Sub-Disciplines

Computer Science

SUBJECT CLASSIFICATION

Natural Language Processing

TYPE (METHOD/APPROACH)

Research work on speech processing for Telugu language

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 11, No.2

editor@cirworld.com

www.cirworld.com, member.cirworld.com

ISSN 2277-3061

1 INTRODUCTION

The objective of the text processing component [1, 2] is to process the given input text and produce the written form (orthographic form) of the text into the spoken form. This orthographic form is realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus. For natural sounding speech synthesis [3, 4], it is essential that the text processing component produce an appropriate sequence of orthographic units corresponding to an arbitrary input text. The input to a TTS system is the raw text [6, 7] as available in news websites, blogs, documents etc which contain the required text in font-encodings, native scripts and non-standard words [5] such as addresses, numbers, currency etc. The majority of the issues are associated in building a TTS for a new language is associated with handling of real-world text. Current state-of-art TTS system which are existing for well researched languages like English and other popular languages use rich set of linguistic resources such as word-sense disambiguation [8], morphological analyzer [9], Part-of-Speech tagging, letter-to-sound rules, syllabification, stress-patterns in one form or the other to build a text processing component of a TTS system. Processing the given text to readable [10] form consists in expanding abbreviations, converting names, numbers, acronyms, dates etc. into their spoken form. In real text, many non-standard representations of words appear, for e.g., numbers, abbreviations, acronyms, currency, dates, URLs. All these non-standard representations must typically be normalized to standard words before synthesis. For example, number **30** has to be expanded to "**muppai**" **ముప్పై a cardinal number** or to "**muppai rUpAyalu**" , **ముప్పై** రుపాయలు or **muppai nimishalu**" , **ముప్పై** నిమిషాలు **.**

This paper, presents the ongoing efforts made to identify the various non-standard representations of the words in unrestricted real Telugu text and ways of converting them to standard words. A novel approach to text processing is proposed, wherein the first level we combine tokenization and initial token classification as one stage, followed by a second level of token sense disambiguation. Finally the standard word representation is achieved using the expansion rules and the look up table (database).

1.1 Nature and Format of Telugu Text

Telugu is a South-Central Dravidian language predominantly spoken in the South Indian state of Andhra Pradesh where it is an official language. One of the four classical languages of India, Telugu ranks third by the number of native speakers in India (74 million), thirteenth in the Ethnologue list of most-spoken languages worldwide. It is one of the twenty-two scheduled languages of the Republic of India. Telugu has several features of Sanskrit that have subsequently been lost in Sanskrit's daughter languages such as Hindi and Bengali, especially in the pronunciation of some vowels and consonants. The scripts in Indian languages are stored in digital computers in ISCII, UNICODE and in transliteration schemes of various fonts. The input text could be any of these formats.

2 PROPOSED MODEL FOR TEXT NORMALIZATION

The architecture of text normalization (see Figure 1) consists of various modules like tokenization, token classification, token sense disambiguation and standard word generation. The proposed system is designed with a set of hand crafted rules which are required at each phase. The modularization is made keeping in view the nature of Telugu language. These modules convert the non standard representations of words in the sentences into standard representations by applying the rules at each phase. The following description gives the detailed explanation of each module.

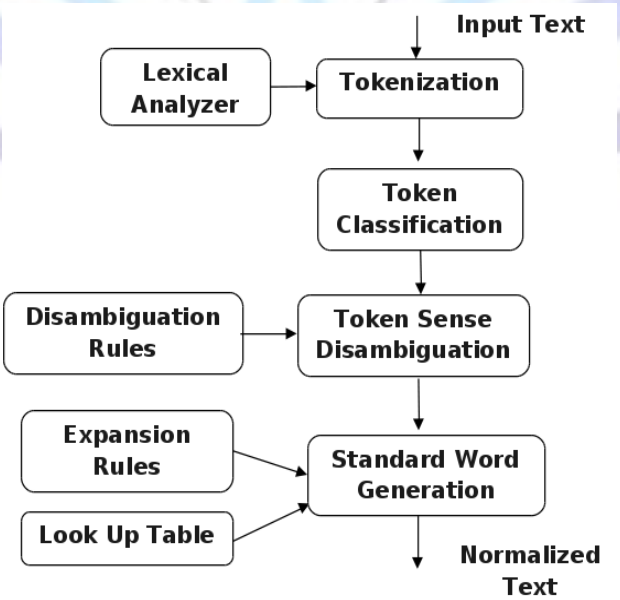

Fig 1: Architecture for Text Normalization

2.1 Tokenization and Token Classification

A token is a string of characters, categorized according to the rules as a symbol. Tokens are defined by regular expressions, which are understood by a lexical analyzer generator. Tokenization and token classification are performed using a lexical analyzer. Tokenization is the process of breaking up a stream of text into words, phrases, symbols, or other meaningful elements called tokens. Each token is separated by spaces or entries in the separation list and is classified using the various classifications.

Initially an input text is taken and it is chunked into sentences based on the sentence delimiter. Whitespace is the most commonly used delimiter between words and is extensively employed for tokenization. But using whitespace as the only delimiter has a number of shortcomings: a token type which allows the occurrence of whitespace within the token will not be recognized as a single token, but split up into two or more tokens. Further, an important shortcoming is that every token so obtained will then have to go through a token identification process that identifies its token type/category. This approach might not be feasible for some languages like Chinese and Japanese which do not use any form of white space between words. In the proposed approach, tokenization and initial token classification/identification are achieved in a single step.

2.2 Token Sense Disambiguation

Once the tokens are extracted from the input text, the category of each token is identified by the lexical analyzer itself when there is no ambiguity arising from the formats of token. In case of ambiguity, the token with the possible token categories is given as output to facilitate further disambiguation. Disambiguation is generally handled by hand-crafted context-dependent rules which are difficult to write, maintain and adapt to new domain. Identification of token category involves high degree of ambiguity. The process of resolving the conflicts that arise when a single term is ambiguous is called token sense disambiguation. These conflicts are resolved using disambiguation rules. Token sense disambiguation can be mapped to general homograph disambiguation using decision-list based data-driven techniques. When dealing with unrestricted text input ambiguity exists among cardinal, ordinal, decimal numbers and time formats, etc. To remove this ambiguity a set of rules are framed and implemented to achieve token sense disambiguation.

2.3 Standard Word Generation

Standard word generation is accomplished by using expansion rules and look-up table. After identifying and disambiguating the token categories, expansion of non standard words is done by a combination of rules (e.g., for expanding numbers, currency, dates) and look-up tables (e.g., for abbreviations and acronyms). Token-To-Word rules are written, which are specific for each token type, and for each format within a token type. Finally all the non-standard representations are converted to standard words, or normalized which would then be processed in various applications.

3 IMPLEMENTATION OF THE SYSTEM

To explain in detail about the working of all the modules in the system an input text is taken as given below which includes various categories of input token types like time, money and normal string literals.

మన రాష్ట్ర ప్రభుత్వం ఉగాది రోజున సా. 5:30 వరకు రైతు బజార్ లో ప్రతీ 50రూ ఖరీదు పై 4.50 తగ్గించింది.

3.1 Tokenization

The given input will be divided into tokens by using the lexical analyzer. White space delimiter is used to tokenize the words of the text input.

మన, రాష్ట్ర, ప్రభుత్వం, ఉగాది, రోజున, సా., 5:30, వరకు, రైతు, బజార్, లో, ప్రతీ, 50రూ, ఖరీదు, పై, 4.50, తగ్గించింది.

3.2 Token Classification

In this phase, the tokens are classified based on their types and the tokens that are in the standard form does not require any kind of classification. Only సా., 5:30, 50రూ, 4.50 tokens are classified as they are not in standard form. The first two tokens (సా., 5:30) are given to time group, third token (50రూ,) is given to currency, and fourth token (4.50) is given to decimal numbers.

3.3 Token Sense Disambiguation

In the first three tokens, there is no ambiguity. So they are passed to next phase. Whereas the fourth token *4.50* in the above phase is considered as a decimal number and the expected output will be *four point five zero*, but the context in which it is used represents money. So the output should be in the form *four rupees fifty paisa*. Thus there is a necessity to frame the disambiguation rules to avoid ambiguity when classifying the tokens.

3.4 Standard Word Generation

During this phase the abbreviated tokens will refer to the look up table and the remaining tokens will refer to the

ISSN 2277-3061

expansion rules, based on this normalized text is generated.

సా. - సాయంత్రం

5:30 - ఐదు గంటలు ముప్పై నిమిషాలు

50రూ - యాభై రూపాయల

4.50 – నాలుగు రూపాయల యాభై పైసలు

3.5 Normalized Output

మన రాష్ట్ర ప్రభుత్వం ఉగాది రోజున సాయంత్రం ఐదు గంటలు ముప్పై నిమిషాలు వరకు రైతు బజార్ లో ప్రతీ యాభై రూపాయల ఖరీదు పై నాలుగు రూపాయలు యాభై పైసలు తగ్గించింది

4 RESULTS

4.1 Numeric

4.1.1 Cardinal Numbers

A Cardinal Number is a number that gives the count of something, such as one, two, three, four, five. Cardinal numbers are also known as "counting numbers," because they represent the quantity**.** Cardinal numbers (Table 1) may appear in the input text to give the count of any unit like population, money, etc,.

Table 1. Cardinal Numbers

Digit	English Conversion	Telugu Conversion
3	Three	మూడు
24	Twenty-four	ఇరవై నాల్గు
100	Hundred	వంద
1000	Thousand	వెయ్యి
10000	Ten thousand	పది వేలు

4.1.2 Ordinal Numbers

An Ordinal Number is a number that tells rank or the position of something in a list. 1st, 2nd, 3rd, 4th, 5th etc. When an ordinal number (Table 2) occurs it should not be spoken as a cardinal number, for example number **“3”** should be spoken as **“third”** but not **“three”**.

Table 2. Ordinal Numbers

Digit	English Conversion	Telugu Conversion
1	First	మొదటి
5	Fifth	ఐదవ
23	Twenty-third	ఇరవై మూడవ

4.1.3 Decimal Numbers

Decimal numbers consist of a point or dot which separates the whole number part from the fractional part of the number. In this kind of inputs there is a possibility of ambiguity due to the context in which the decimal number is used. When dealing with decimal numbers the fractional part need to be expanded differently depending on the context it is used. Few examples (Table 3) are mentioned which lead to word sense disambiguation depending on the context it is used.

ISSN 2277-3061

Table 3. Decimal Numbers

Input Text	Normalized Output
5.68	ఐదు పాయింట్ ఆరు ఎనిమిది
6.40 ఖరీదు	ఆరు రూపాయలు నలభై పైసలు
2.5 టేబుల్ స్పూన్లు	రెండున్నర టేబుల్ స్పూన్లు
9.25 లీటర్ల	తొమ్మిదింపావు లీటర్ల

4.1.4 Phone Numbers

A telephone number or phone number is a sequence of digits. Here each individual digit needs to be extracted and its number name has to be spoken rather than treating it as a number that represents currency or ordinal or cardinal number. An example is shown in Table 4.

Table 4. Telephone Numbers

Input Text	Normalized Output
నెం 9433076654	నంబర్ తొమ్మిది నాలుగు మూడు మూడు సున్నా ఏడు ఆరు ఆరు ఐదు నాలుగు

4.2 Date Formats

There are various date formats like dd/mm/yy, mm/dd/yy, yy/dd/mm where dd represents day, mm represents month and y/yy represents year. Day, month and year can be delimited by either using '/' or '-'. In Telugu language (Table 5) the date can be give as month name followed by day and year.

Table 5. Date Formats

Input Text	Normalized Output
23-2-2012 23/2/2012	ఇరవై మూడు ఫిబ్రవరి రెండు వేల పన్నెండవ సంవత్సరం
Oct 15, 2012	అక్టోబరు పదిహేను రెండు వేల పన్నెండవ సంవత్సరం

4.3 Time

Telugu language uses a set of representations (Table 6) to mention the time depending on whether it is morning/evening or day/night. Time of a day is written in the 24-hour notation or 12-hour notation in the form hh:mm (for example 01:23) or hh:mm:ss (for example, 01:23:45), where hh is the number of hours, mm is the number of minutes and ss is the number of seconds.

Table 6. Time Formats

Input Text	Normalized Output
Standard Time Format	hh:mm:ss
Time Separator	Colon(:)
ఉ.	ఉదయం
మ.	మధ్యాహ్నం
తె.	తెల్లవారుజాము
రా.	రాత్రి

సా.	సాయంత్రం
5:30:15	ఐదు గంటలు ముప్పై నిమిషాలు పదిహేను సెకన్లు
9:20	తొమ్మిది గంటలు ఇరవై నిమిషాలు
6 గంటలకి	ఆరు గంటలకి

4.4 Currency

The representation of currency in Telugu is given by using **"రూ", any** numeric value followed by this symbol indicates that the given numeric value is rupees and therefore it should be expanded using the currency notation. In Telugu language (Table 7) we first speak out the numeric value followed by the word **"రూపాయలు"**. Here the expansion should not be like phone numbers it should contain the currency notations like crores, lakhs, thousands, etc.

Table 7. Currency

Input Text	Normalized Output
రూ	రూపాయలు
523రూ	ఐదు వందల ఇరవై మూడు రూపాయలు

4.5 Abbreviations and Acronyms

An abbreviation is a shortening by any method; like taking a letter or group of letters taken from a word or phrase. An acronym is an abbreviation formed from the initial components in a phrase or word. Few examples (Table 8) are mentioned where, in the real text we find "Dr" but we speak out as "Doctor".

Table 8. Abbreviations & Acronyms

Input Text	Normalized Output
యునిసెఫ్ (UNICEF)	యునైటెడ్ నేషన్స్ ఇంటర్నేషనల్ చిల్డ్రన్స్ ఎమర్జెన్సి ఫండ్
తెదేపా(TDP)	తెలుగు దేశం పార్టీ
తెరాస(TRS)	తెలంగాణ రాష్ట్ర సమితి
భాజపా(BJP)	భారతీయ జనతా పార్టీ
Dr.,డా.	డాక్టర్
St.,సె.	సెయింట్
Kg.,కే.జి.	కిలో గ్రామ్
Mm,మి.మి.	మిల్లిమీటర్

ISSN 2277-3061

4.6 Address

The address consists of text which is to be read as it appears and few numbers like house number or flat number or a pin code which need to read accordingly. The expansion of house number or flat number is done using money notation whereas the pin code is represented as individual digits like phone number.

Table 9. Address

Input Text	Normalized Output
ఇం.నెం:43-456/ఈ, న్యూ బాకారం, హైదరాబాదు ,500037.	ఇంటి నెంబరు నలభైమూడు డాష్ నాలుగువందల యాభై ఆరు బై ఈ న్యూ బాకారం హైదరాబాదు ఐదు సున్న సున్న సున్న మూడు ఏడూ

4.7 Others

4.7.1 Percentages

A percentage is a number or ratio as a fraction of 100. It is denoted using the percent sign, "%". A percentage can be a single number or can range between two numbers, which is represented using "-". Then the input text containing (Table 10) such numeric's like **"20 – 30%"** should be read out as **"twenty to thirty percent"** not as **"twenty minus thirty"**.

Table 10. Pecentage

Input Text	Normalized Output
20-30%	ఇరవై నుంచి ముప్పై శాతం వరకు

4.7.2 Symbols

When preprocessing the input text many symbols may appear which may need to be represented accordingly. Special care has to be taken to handle the different symbols (Table 11) that can be seen during text processing.

Table 11. Symbols

Input Text	Normalized Output
$	డాలర్
£	యూరో
%	శాతం
:	నిశ్పత్తి

4.8 Coverage Analysis

The performance of the system (Table 12) is tested on the data obtained from different sources (web). The data is collected from various domains like news articles, sports, stock market related information, politics, business, health and etc. The test data was manually prepared in such a way that it consisted of all categories of tokens. The data collected consisted of 3250 lines with 28600 words.

Table 12. Coverage Analysis

NSW Category	No. of Occurrences	Prediction Accuracy
Cardinal Numbers	1348	99.55%
Ordinal Numbers	713	99.71%
Float Numbers	1209	98.17%
Date	503	98.20%
Time	242	98.34%
Abbreviations & Acronyms	202	98.51%
Symbols	123	97.50%
Websites & URLs	100	100%
Address & Contact Numbers	136	99.79%

5 CONCLUSIONS

This paper presents the need for text to be preprocessed before it is handed to any synthesizer. The real unrestricted text may contain many different non standard formats, which need to be converted to a standard representation. The token types handled are Numbers (cardinal, ordinal and decimal), Phone Numbers, Time, Date and Year, Symbols, URL's, Websites, Addresses and Literal Strings. A rule based system is designed to achieve word sense disambiguation and the standard word representation. An effort is made to identify most of the non-standard words that can occur in the unrestricted text and rules are designed, implemented and tested to a great extent. However for languages like Telugu, where research is in progress does not have enough linguistic resources, it involves several complexities starting from accumulation of text corpora in digital and processable format. Linguistic components are not available in such rich fashion for all languages of the world.

6 REFERENCES

[1] Anand Arokia Raj, Tanuja Sarkar, Sathish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad, Alan W Black," Text Processing for Text to Speech Systems in Indian Languages " in in Proceedings of 6th ISCA Speech Synthesis Workshop SSW6, Bonn, Germany, 2007. Report No: IIT/TR/2007/32

[2] A. Acero, H. Hon, and X. Huang,

Spoken Language Processing: A guide to Theory, Algorithm, and System Development, Prentice Hall PTR, 2001.

[3]Speech communications Human &Machine by Douglas O'Shaughnessy

[4] Speech and Language Processing by Daniel Jurafsky & James H.Martin

[5]Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf and Christopher Richards. Normalization of Non-StandardWords. *Computer Speech and Language*, 15(3):287.333, 2001.

[6]Jonathan Allen, M. Sharon Hunnicutt, and Dennis Klatt. *From Text to Speech: The MITalk System* Cambridge University Press, Cambridge, 1987.

[7]Linguistic Data Consortium: Text Conditioning Tools http://morph.ldc.upenn.edu/ Catalog/LDC98T31.html, 1998

[8] Yarowsky, David. Homograph Disambiguation in Text-to-Speech Synthesis In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, Springer, New York, pages 157.172, 1996.

[9]Taylor, P., R. Caley, and A.W. Black. The Edinburgh Speech Tools Library http://www.cstr.ed.ac.uk/projects/speechtools.html, 2002

[10]Published a paper "Bhaashika-Telugu TTS system", in the Journal IJEST- International Journal of Engineering Science and Technology, Vol. 02, Issue 11,Nov 2010, ISSN: 0975–5462.

AUTHOR PROFILE

Dr.K.V.N.Sunitha Currently working as Principal, BVRIT Hyderabad college of Engineering for women, Nizampet, Hyderabad has done her B.Tech ECE from NagarjunaUniversity, M.Tech Computer Science from REC Warangal. She completed her Ph.D from JNTU, Hyderabad in 2006. She has 21 years of Teaching Experience, worked at various engineering colleges. She received "Academic Excellence Award" by the management of G.Narayanamma Institute of Technology & Science on 18th September 2005. She also received "Best computer Science engineering Teacher award for the year 2007" by Indian Society for Technical Education ISTE. She has been recognized & invited by AICTE as NBA expert evaluator. Her autobiography was included in "Marquis Who is Who in the World " , 28th edition 2011, since August 2012. She has authored four text books, "Programming in UNIX and Compiler design"- BS Publications & "Formal Languages and Automta Theory" by Tata Mc Graw Hill , " Theory of Computation" by TMH in 2011, “Compiler Construction by Pearson India pvt ltd. She is an academic advisory member & Board of Studies member for other Engineering Colleges. She has published more than 75 papers in International & National Journals and conferences. She is a reviewer for many national and International Journals. She is fellow of Institute of engineers, Sr member for IEEE & International association CSIT, and life member of many technical associations like CSI and ACM.

Mrs.P.Sunitha Devi presently working as Assistant Professor in CSE Dept, G.Narayanamma Institute of Tech& Science, Hyderabad. She has completed her M.Tech from JNTU Hyderabad and currently pursuing Research in the area of Telugu Text to Speech Synthesis. She is Life Member of CSI.