Kannada Text Normalization in Source Analysis Phase of Machine Translation System

Prathibha R J [#1], Padma M C [*2]

[#] Department of Information Science and Engineering,
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India.
[1] rjprathibha@sjce.ac.in

[*] Department of Computer science and Engineering,
P E S College of Engineering, Mandya, Karnataka, India.
[2] padmapes@gmail.com

***Abstract*— Almost all documents used in text processing applications contain raw or real text. Some of words in raw text are represented in non-standard form. In this context, there is a need of text normalizer to transform or convert non-standard forms of words into standard and consistent forms. Design of text normalizer depends on the kind of data and applications. In Machine Translation System (MTS), a normalizer is required to categorize raw input text into morpheme based and non-morpheme based words and process non-morpheme based words by assigning their respective Parts of Speech (PoS) tags. In this paper, a text normalizer is proposed to normalize Kannada source text in MTS. The proposed text normalizer is tested on Enabling Minority Language Engineering (EMILLE) corpus and nearly 45%-57% of input text has been filtered during normalization itself.**

Keyword - Machine Translation System, Source Analysis, Text Normalization, Tokenization.

I. INTRODUCTION

Speech and language processing applications require handling of raw text which is in unstructured form. In raw text, many words are represented in non-standard form. These words are called Non-Standard Words (NSW). Few examples for NSWs are acronyms, abbreviations, dates, numbers (time, year, time, floating point, cardinal, ordinal), etc., These NSWs are need to be converted into their standard word forms. Later, these converted texts are processed in various language and speech processing applications like MTS, text-to-speech conversion system, automatic speech recognition system, etc.

In general, text normalization is the process of converting raw text into convenient, consistent and standard form, depending on the type of data and application. For example, in cell phone, users send text in short forms called Short Messaging Service (SMS) text. These types of words are called non-standard words. However, in the analysis of SMS text, pre-translation normalization is required to transform SMS texts into standard form. Hence, text normalization is a prerequisite for a variety of speech and language processing tasks.

In linguistic, morpheme is the smallest unit which carries meaning. For example, the word "going" comprises of two morphemes, "go" and "ing". The word "go" is an individual unit which carries meaning and it cannot be broken further into smaller unit of meaning, hence it is a morpheme.

Kannada is a derivational, inflectional and morphologically rich language. Inflectional or declinable words are generated by adding a set of suffixes to root/stem word. For example, in English language, different inflectional words formed from verb-root "save" with suffixes "s", "ing" and "d" are given below.

- save + s = saves
- save + ing = saving
- save + d = saved

Non-inflectional or indeclinable word is a word that cannot be inflected and remains in the same form for all genders, numbers and cases. For example, words "therefore", "but", "and" are non-inflectional words [1].

Morpheme based words are the words having morphemes, morphological grammatical features associated with them. The morphological features of a word are prefix, stem, and suffix. For example, in English, the word "recompilation", contains the morphemes "re", "compil" and "ation" as prefix, stem and suffix respectively. The grammatical information of a word is case, gender, person, number, tense, etc. Non-morpheme based words are the words having no morphemes with them. For example, acronyms, abbreviations, punctuation-marks, expressions and numerical data (numbers, date, time, Internet protocol address etc.,) are non-morpheme based words. In MTS, input words are identified and classified as morpheme based and non-morpheme based words.

The aim of MTS is to convert input text from one language called source language to target language. Mainly, there are three phases in MTS viz., i) Source analysis phase ii) Semantic analysis phase iii) Target language generation phase. In source analysis phase, input raw texts need to be normalized. In general, raw text contains set of paragraphs. These paragraphs need to be split into sentences and further these sentences into words/tokens. Some of tokens like punctuation marks, numbers, acronyms, abbreviations, etc., that are present in raw text need to be extracted and processed during normalization process itself. In this context, a text normalizer is proposed to normalize Kannada text (source language) in source analysis phase of machine translation system.

The paper is organized as follows. Section II gives the literature survey on existing tokneizer tools and text normalizers. Section III describes the details of proposed text normalizer for Kannada language in machine translation system. In Section IV, performance evaluation and result analysis of proposed text normalizer on EMILEE corpus is explained. Conclusion is given in Section V.

II. LITERATURE SURVEY

Text segmentation and tokenization are two important tasks in normalization of given raw text. In literature, many tokenizers are reported for both Indian and non-Indian languages. But, most of these tokenizers consider space as delimiter and split given text into set of tokens [2-8]. Some of existing tokenizer tools are listed below.

- Word tokenization with python NLTK [2]
- Nipdotnet Tokenizer [3]
- Mila Tokenizer [4]
- NLTK word tokenizer [5]
- TextBlob word tokenizer [6]
- MBSP word tokenizer [7]
- Pattern word tokenizer [8]

These tokenizers work well for both Indian and non-Indian languages. A special, Indic tokenizer [9] is designed specifically for Indian languages. Some limitations are observed in Indic tokenizer. These limitations are listed below.

- Numbers with period, comma, and hyphen split into separate tokens.
- Abbreviations and acronyms are separated based on period (.) as delimiter.
- Digit followed by alphabets or alphabets followed by digits will not be split into separate tokens.

Detailed descriptions about these limitations with sample examples are given in Table I.

Literature shows that existing text normalizers are designed for both Indian and non-Indian languages. But most of these text normalizers are designed for text to speech synthesis applications [10-15]. In literature no text normalizer for MTS is reported. In this context, there is a great demand for the design of text normalizer in MTS. In this paper, text tokenizer and normalizer for Kannada language in MTS are presented.

TABLE I: Few Sample Tokens Obtained by Indic Tokenizer

Input word	Tokens obtained by Indic tokenizer	Expected tokens in MTS
150-200	150-200	150-200
24,000	24,000	24,000
24,500rinda	24,500rinda	24,000rinda
30randu	30randu	30randu
pu.thi.no.	Pu Thi.No.	pu.thi.no.
shee.90rashthu	Shee.90rashthu	Shee.90rashthu
Mr.Prasad	Mr.Prasad	Mr.Prasad
3.6laksha ru.	3.6lakshaRu.	3.6lakshaRu.

III. PROPOSED WORK

Table I shows the limitations of Indic tokenizer with examples. To overcome these limitations, a special tokenizer is proposed. It is also found that in literature, almost all existing text normalizers are specifically designed for text to speech synthesis. Hence a text normalizer for Kannada language in MTS is also proposed.

The architecture of proposed text normalizer in source analysis phase of MTS is shown in Fig. 1. There are six phases in text normalization process, viz., i) Segmentation of text into set of sentences ii) Splitting of sentences into set of tokens, iii) Assignment of unique identification numbers to each token, iv) Identification and classification of tokens, v) PoS tagging for non-morpheme based words, vi) Removal of redundant morpheme based words. Detailed description of these six phases is given below.

i) **Segmentation of text in to set of sentences:** Sentence segmentation is the process of dividing running text into sentences. In natural language processing applications, sentence boundary disambiguation is the major problem to decide where sentences begin and end. Due to the use of full stop character in abbreviations, acronyms, decimal point, email address, etc., may or may not also terminate a sentence. For example, the sentence "Mr. Nuthan went to market.", can be split into two sentences as i) "Mr" and ii) "Nuthan went to market", by considering full stop character as delimiter. By considering such kind of ambiguities, a rule based sentence segmentation tool is proposed.

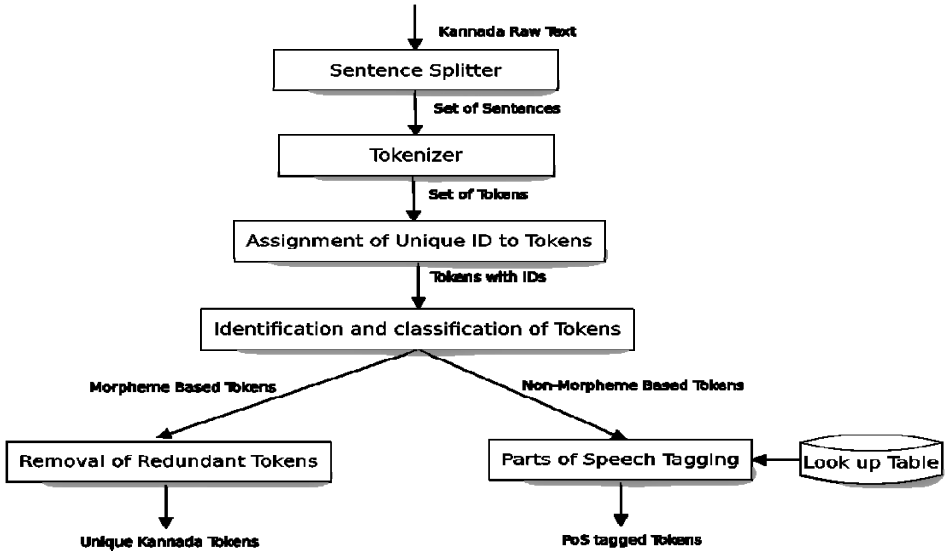

Fig. 1. Architecture of the Proposed Text Normalizer in Source Analysis Phase of Machine Translation System

ii) **Splitting of sentences into set of tokens:** Tokenization is the process of splitting the given sentences into units called tokens. The tokens may be numbers, special symbols, words, punctuation-marks, etc. The proposed tokenizer is designed to overcome the limitations present in existing Indic tokenizer.

iii) **Assignment of unique identification numbers to each token:** For each token in the raw text, an unique identification number is generated and assigned.

iv) **Identification and classification of tokens:** In source analysis phase of MTS, only morpheme based words need to be processed. However, non-morpheme based words are to be tagged with their respective lexical category. In this context, tokens are identified and categorized into morpheme based and non-morpheme based words.

v) **Parts of speech tagging for non-morpheme based words:** In text normalization, non-morpheme based words are handled by assigning PoS tags. The different types of non-morpheme based words and their respective PoS tag notations used in PoS tag set are shown below.

i. ACRO: Acronym
ii. ABBR: Abbreviation
iii. PUNCT: Punctuation
iv. NUMB: Number

A look up table is manually created to store punctuation-marks, abbreviations, acronyms with their relevant PoS tags. Almost all Kannada abbreviations, acronyms and their respective PoS tags stored in look up table. If input text contains Kannada abbreviations or acronyms or punctuation marks, they are searched in the look up table and tagged with their relevant tags ABBR, ACRO, PUNCT respectively. The tokens that contain numbers are tagged as NUMB.

vi) **Removal of redundant morpheme based words:** Morpheme based words' list may contain redundant words. Analysis of only one occurrence of such redundant words is sufficient. Hence, redundant words are removed from morpheme based word list.

IV. PERFORMANCE EVALUATION AND RESULT ANALYSIS

Publicly, no standard Kannada data set is available for research purpose. However, the EMILLE corpus is distributed free of cost for use in non-profit-making research. We have chosen 50 and 25 documents from stories and novels category of EMILEE corpus. These documents contain punctuation-marks, numbers, special symbols, words, acronyms, abbreviations in Kannada. The result obtained by proposed text normalizer on chosen Kannada EMILEE corpus is shown in Table II. The performance evaluation of proposed text normalizer is calculated using the following formulae.

Total number of unique Kannada words (T_u) = $T_m - T_r$ (1)
Percentage of tokens filtered during normalization (T_f) = $T_u / T_t * 100$ (2)
Where
T_u – Total number of Kannada unique words
T_m – Total number of morpheme based Kannada words
T_r – Total number of redundant morpheme based Kannada words
T_t – Total number of tokens obtained by the proposed tokenizer

TABLE II Results Obtained by the Proposed Text Normalizer on Kannada EMILEE Corpus

Document Type	**No. of Documents**	**No. of tokens from Indic tokenizer**	**No. of Tokens from proposed tokenizer (T_t)**	**No. of non-morpheme based words (T_n)**	**No. of morpheme based words (T_m)**	**No. of Redundant words (T_r)**	**No .of Unique words (T_u)**	**% ge of Tokens filtered (T_f)**
Stories	50	45423	44269	8180	37243	17129	20144	45.50
Novels	25	32587	29438	7225	22213	5433	16780	57.00

It is observed from the Table II is that the content of input Kannada raw text is processed and normalized. During normalization, tokens are categorized into morpheme based (37243 and 22213 words) and non-morpheme based tokens (8180 and 7225 words). All non-morpheme based words are tagged with their respective PoS tags. The redundant words are removed from morpheme based words' list (17129 and 5433 words). Hence nearly 45%-57% of input tokens are processed and filtered during normalization process itself.

V. CONCLUSION

Input for almost all natural language and speech processing applications are in the form of raw or real text. Hence normalization of non-standard form of words is very much essential. In machine translation system, the main objectives of normalization is to tokenize the input raw text, identify and classify tokens into morpheme based and non-morpheme based words, PoS tagging of non-morpheme based words by using lookup table and removal of redundant words in morpheme based words' list. The proposed Kannada text normalizer is tested on EMILEE corpus. Nearly 45% - 57% of input text has been processed and filtered. The remaining 43% - 55% of input text are morpheme based words. These words are further processed in machine translation system.

REFERENCES

[1] R.J. Prathibha and M.C. Padma, Design of Rule based Lemmatizer for Kannada Inflectional Words, International Conference on Emerging Research in Electronics, Computer Science and Technology, 2015, p. 264-269.
[2] http://nlpdotnet.com/services/Tokenizer.aspx
[3] http://www.mila.cs.technion.ac.il/tools_token.html
[4] http://textanalysisonline.com/nltk-word-tokenize
[5] http://textanalysisonline.com/textblob-word-tokenize
[6] http://textanalysisonline.com/mbsp-word-tokenize
[7] http://textanalysisonline.com/pattern-word-tokenize
[8] http://text-processing.com/demo/tokenize
[9] http://github.com/anoopkunchukuttan/indic_nlp_library
[10] Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf and Christopher Richards, Normalization of non-standard words, Computer Speech and Language, 2001 p. 287-333.
[11] Gerasimos Xydas, Georgios Karberis, and Georgios Kouroupertroglou, Text Normalization for the Pronunciation of Non-standard Words in an Inflected Language,© Springer-Verlag Berlin Heidelberg 2004, p. 390-399.
[12] AiTi Aw, Min Zhang, Juan Xiao, Jian Su, A Phrase-based Statistical Model for SMS Text Normalization, Proceedings of the COLING/Association for Computational Linguistics (ACL) main Conference Poster Sessions, 2006, p. 33-40.
[13] Paul Cook and Suzanne Stevenson, An Unsupervised Model for Text Message Normalization, Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, 2009, p. 71-78,
[14] Jagadish S Kallimani, Srinivasa K G, Eswara Reddy B, Normalization of Non Standard Words for Kannada Speech Synthesis, International Journal of Advances in Computer Science and Technology, 2012, p. 21-26.
[15] Deana L. Pennell, Yang Liu, Normalization of informal text, Computer Speech and Language-Elsevier Publications, 2014, p. 256-277.

AUTHOR PROFILE

R. J. PRATHIBHA received her B.E. in Computer Science and M. Tech. in Software Engineering from Visvesvaraya Technological University (VTU), Belgaum, India. Currently, she is working as Assistant Professor in the department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India. Her research areas are Natural Language Processing, Machine Translation, Artificial Intelligence, Machine Learning, Data Mining and warehousing and Big Data Analytics. She has published 10 papers in International Conferences/Journals

M. C. PADMA received B.E. and M. Sc. Tech. by Research degree in Computer Science and Engineering from University of Mysore, India, and Ph.D. degree from Visvesvaraya Technological University, Belgaum, India. Currently, she is Professor and Head, Department of Computer Science and Engineering, PES College of Engineering, Mandya, India. She is a member of IEEE, MISTE, CSI, IEI professional societies. Her research areas are Pattern Recognition, Natural Language Processing, Document Image Analysis and Recognition. She has published 45 papers in National/International Conferences/Journals and conducted National/International Conferences and Editor for Proceedings of International Conference Emerging Research in Electronics, Computer Science and Technology.