CSE6060 - SNLP

Task 4: Review of 3 research papers of NLP used for Regional Languages

19MAI0002

---

1. A Survey of Automatic Text Summarization System for Different Regional Language in India

   Main inspiration of text summarization system is to give short insight but same meaningful context of large original documents. Extractive summarization is like highlighting the document i.e. important sentences according to ranking are taken for summary. The abstractive summarization is done by complete understanding the whole document through NLP. First authors studied different summarizer used for Indian languages like Hindi, Bengali, Punjabi, Tamil, Kannada, Urdu and did the performance comparison among them. Authors proposed Extractive Text Summarization for Marathi Language here. Data can be used for this summarizer are single stories, single or multiple news documents. Data pre-processing involves removing stop wards, stemming and portioning document into a collection of sentences. Stemming is done by comparing each word with Marathi morph if not present then word will be stemmed and checked again. Words again normalised if they are similar and sentences also deleted if the are duplicated. Marathi Keywords are identified by calculating TF-ISF (Term Frequency-Inverse Sentence Frequency) calculation. Mostly Numeric data is included in summary so as the headlines from news articles. Proper names used for person, place, concept is identified by Marathi Named Entity Recognition. Based on thematic term and positions are ranked for extractive summarization. Cue phrases in English like ‘in conclusion’, ‘finally’ which has high importance also applied for Marathi summarizer by firstly creating their list and sentences using them will be included in summary.

2. Marathi Text-To-Speech Synthesis using Natural Language Processing

   Considering high use Text-TO-Speech technology use in various areas authors designed model for Marathi language spoken many people in 3rd largest state in India. This TTS synthesis system based on formant and concatenation techniques. Marathi phonetic system made up 43 consonants and 28 vowels and has limited syllabic structure. Pre-processing includes detecting spaces, sentences, punctuations, words and non-Marathi words. During language processing authors have maintained Exception dictionary database whose pronunciations given explicitly and Marathi Letter to Sound rules Database whose pronunciations made according rules. After passing deciding stress pattern word’s pronunciation has checked with Exception database if matched then directly given to pronunciation. And if not then flag added used as suffix or prefix from Exception database are considered, means they are applied with pronunciation rules. Thus, checked with another database and after removing prefixes/suffixes they are passed for pronunciation. Rules based database suggest best pronunciation

according best matching rule like followed by, preceded by. Group of words may affect pronunciation, so they are enclosed in braces and stress pattern worked accordingly. Marathi phonemes has all the data for production of sound. Formant synthesis used for vowels and concatenation synthesis used for special Marathi consonants. Exception dictionary has rules according different lexical stress levels. Digital signal processing responsible pronunciation output works with phonemic representation and pause flags and generate output in wave file. Authors able to achieve 91% accuracy for letter-to-sound system. According to 10 volunteers' observation they able to receive 'good' remark.

3. ANALYSIS OF MWES IN HINDI TEXT USING NLTK

Authors tried pre-processed Hindi words using repositories provided by NLTK and worked on analysis of Multi Word Expressions. Hindi novel 'KarmaBhumi' by Munshi PremChand is used as Hindi corpus and analysed according corpus by CFILT (Centre for Indian Language Technology Solutions). As MWEs are combination of words which together mean differently than individual. Consider example: बाल ववाह (Baal Vivah). Dataset a novel कम भूमी (Karama Bhumi) has 14081 unique words helps for better corpus divided in 5 sections out of which last 4 sections used for training. For data pre-processing new corpus is created using NLTK and added to python directory. NLTK is enough cable to perform different functionalities on lots of languages. For figuring out the Multi-words, Bi-grams, Tri-grams, n-grams are extracted from text. Considering existing types and generating new types authors made list of different types of Hindi MWEs: Acronyms and Abbreviations, Complex Predicates, Foreign Words and Terms, Idioms, Morphemes, Replicating words, Proverbs, Named Entities, Adverbs and their further specific types.
The classification tools used for the Hindi MWEs is Protégé, which provides ontological
classification for the various class types. F-score is calculated by performing experiments with Hindi corpus by CFILT where they get excellent 90+% accuracy for acronyms and abbreviations, Replicating class, 'vaala class, compound adverbs but faced difficulties in identifying with acronyms and named entites.

References:

1. Giri, V. V., Math, M. M., & Kulkarni, U. P. (2016). A Survey of Automatic Text Summarization System for Different Regional Language in India. Bonfring International Journal of Software Engineering and Soft Computing, 6(Special Issue Special Issue on Advances in Computer Science and Engineering and Workshop on Big Data Analytics Editors: Dr. SB Kulkarni, Dr. UP Kulkarni, Dr. SM Joshi and JV Vadavi), 52-57.
2. Kayte, S., Mundada, M., & Kayte, D. C. (2015). Marathi text-to-speech synthesis using natural language processing. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume, 5, 63-67e.
3. Singhal, R. J. A. ANALYSIS OF MWES IN HINDI TEXT USING NLTK.