An efficient tamil text compaction system

Technology, Mobile Phone

An Efficient Tamil Text Compaction System N. M.. Revathi, G. P. Shanthi, Elanchezhiyan. K, T V Geetha, Ranjani Parthasarathi & Madhan Karky Tamil Computing Lab (TaCoLa), College of Engineering Guindy, Anna University, Chennai. haisweety18@gmail. com, jijutodo@gmail. com, madhankarky@gmail. com Abstract Tamil is slowly becoming the online language and mobile text messaging languages for many Tamils around the world. Social networks and mobile platforms now extensively support Unicode and applications for keying Tamil text. The number of characters in a text message is limited in some social nets and mobile text messages. The need for compacting the text becomes essential as it translates to saving online storage space, cost and many more factors. The paper proposes a text compaction system for Tamil, a first of its kind in Tamil. The system proposed in this paper handles common Tamil words, acronyms/abbreviations and numbers. Morphological analyzer [1] and Morphological generator are used to stem inflexion words and replace them to compact using a mapping repository. The proposed work is tested with over 10, 000 words and it is found that the final result is reduced to 40% of the original text. The paper concludes by discussing possible extensions to this system. 1. Introduction: In all languages, using compact or short form of words in text messages, emails, and blogs is rapidly increasing. It is particularly popularly amongst young urbanities as it allows for voiceless communication, useful in noisy environment that would defeat a voice conversation and also buffered communication since the message the sender wants to convey can be accessed by the receiver at any time. Compacting text is thus necessary because of limited message length in blog

sites and tiny user interface of mobile phone. Getting the shortest word has no rule and it is mainly aimed at understanding. That is, those words should be understood by everyone. We can obtain the compact words by omitting letters, replacing prefix and suffix of through suitable symbols and numbers. This causes the compacted system to be credited with creating a language. The paper proposes a Text Compaction system for Tamil, the primogenital in Tamil.. 2. Background: Tamil is perhaps the only classical language, whose glorious literatures date back to the pre-Christian era, has remained in continuous use for more than many millennia now. Due to the untiring efforts of scholars, researches and enthusiasts, it has also evolved creatively over the years to the extent that it is also used today profusely in computers, internet, mobile phone etc. Diverse creative efforts have been taking place that would pave the way for a quantum jump in the usage of Tamil in Information Technology. “ Tamil Virtual University", “ Centre for Research and Applications of Tamil in Internet", 267 “ Tamil Software Development Fund" is to quote a few. These efforts paved the way for the motivation of proposing Tamil compaction system in Tamil. Many compaction systems have been developed for English and other languages. Lee Ming Fung in [2] proposed a Short form Identification and Categorization model based on maximum entropy to identify short forms from actual words and acronyms/abbreviations and categorize the short forms into the short forms formed from letter omission and those formed through phonetic substitution of parts of words. In the proposed system the compact words are formed in a diverse variety of ways such as omission, truncation and phonetic substitution. Acronym Identification and detection has been much

researched. Acrophile in [3] automatically searches acronyms from acronym-expansion pairs from domain specific databases. By acronyms expansion pairs, we refer to a pairs each containing acronyms and their full expanded form or meaning. The paper makes use of acronym expansion pairs to replace the full expanded form with the acronyms. 3. Text Compaction Framework: The figure below presents the various components of the framework. 3. 1 Input Processing The input text is tokenized based on a delimiter and is passed on to the Morphological Analyzer. The analyzer removes the suffix (if present) added to the word and delivers the root word (RW). For example if the input to the analyzer is à®•à®£à®¿ à¯†à®ªà®¾à®±à®¿à®¯à®¿ 3. 2 Identification of the type The proposed paper handles three categories of words; common Tamil words, Abbreviations /acronyms, numbers. Now, the category to which the RW belongs is to be identified. The RW is checked to decide the category of abbreviations /acronyms. This is done by comparing the root word with the keys of the hash map (2. 3). If the comparison results are true then the RW is considered as the abnormal word (AW) i. e. it belongs to the category of acronyms/abbreviations, else, it is treated as the normal word (NW) i. e. it belongs to either the first or third category. the output is given as à®•à®£à®¿ à¯†à®ªà®¾à®±à®¿. 268 3. 3 Extraction of the compact word If the word is identified as a normal word, it is passed to a tree which is built dynamically from the set of words that has already been stored in the dictionary. The NW is then searched in the binary search tree. On finding the NW in the binary search tree, the compact word is retrieved with an efficient mapping algorithm that maps each of the normal word with its compact

word. Say suppose the word is an abnormal word, its compact word is retrieved in the following manner. A linked hash map is built for all the abbreviated words. The hash map uses the first word the abbreviated word as its key. Again with the help of an efficient mapping algorithm, the compact word is retrieved. In case the NW is a number name it is replaced with the numerals based on the place value system. 3. 4 Output Processing The compact word that is being extracted is passed on the Tamil tool Morphological Generator to add the suitable suffix to cater to the rules of the language. 4. Results and Analysis: The paper proposes the following layout for displaying the results to the user. It has two text areas: the one on the left is for entering the input text and the other on the right for displaying the output. The user can also view the no of characters that have been reduced in the output text. Efficiency of the system can be calculated as (no of characters in the input text / no of characters in the output text) X 100%. The proposed work is tested with over 10, 000 words and it is found that the final result is reduced to 40% of the original text. 269 5. Conclusion and Future work: The paper describes the Tamil Compaction System, a framework for shrinking the text such that its meaning remains the same. Different subsystems and components of the framework are described in detail. Results from the implementation of this Tamil compaction system framework is provided and is compared against the compacting third party applications of social networking sites that are implemented for English language. Improving the mapping for words which are frequently used, conceptual reducing, integrating numerical analyser will take this system to its next level. References: Anandan, R. Parthasarathi, and T. V. Geetha,

Morphological Analyser for Tamil. ICON 2002, 2002. Fung, L. M. (2005). SMS short form identification and codec. Unpublished master’s thesis, National University of Singapore, Singapore Acrophile (LSLarkey, P Ogilvie, MA Price, B Tamilio, 2000) a system that automatically searches acronym expansion pairs. Short Message Service (SMS) Texting Symbols: A Functional Analysis of 10, 000 Cellular Phone Text Messages by Robert E. Beasley, Franklin College. 270