Syllabification of Romanized Gujarati Words using Machine Learning

Payal Joshi
Department of Information and Communication Technology,
Veer Narmad South Gujarat University, Surat, India

Abstract- Transliterated search is the current requirement for searching native language contents. To search Gujarati language content, user may enter query in Roman script. This query is then back-transliterated to Gujarati script to search matching documents. To transliterate a Romanized Gujarati word using machine learning model like Condition Random Field, it has to be syllabified first. Word syllabification is done mostly with rule based approach. We have experimented with machine learning approach and achieved significantly more accuracy as compared to rule based approach for Romanized Gujarati word syllabification. In this paper, Maximum Entropy (MaxENT) machine learning model is used for Romanized Gujarati word syllabification.

CCS Concepts- Natural Language Processing, Information Retrieval

Keywords- Romanized Gujarati word syllabification; Information Retrieval

I. INTRODUCTION

Due to digitization, massive contents of a variety of Indian languages is published on the internet. Searching such contents is complicated for the users because of lack of native language input tools. Gujarati is one of the most popular language of India. To search Gujarati contents, user may search using Roman script. E.g. User should be facilitated to type “Surat” to retrieve documents related to Surat in Gujarati language. To back-transliterate Romanized Gujarati word, first it has to be syllabified. Then syllabified word fragments are converted to Gujarati script and combine back to make whole word.

So syllabification is a crucial part of transliterated search. Word syllabification for transliteration is the process of phonetically dividing words into syllables. Syllables are parts that are made up of a vowel sound with or without a closely combined consonant sound.

For word syllabification, rule based approach or machine learning approaches can be used. In this paper, we have used Maximum Entropy (MaxEnt) machine learning model. In results section, we have also shown result of comparison of performance of both the approaches and found MaxEnt performing better.

Basically, maximum entropy model is used for classification in which data is described as a large number of features. MaxEnt can also perform complex NLP tasks like part of speech tagging, sentence detection, word syllabification, and named entity recognition.

In Gujarati language work on syllabification has been done but most of the work is done using rule based approach. Machine learning based approach is required for syllabification so that we can achieve more accuracy.

In this paper, machine learning approach is used for syllabification of Romanized Gujarati word with the intention for more accurate syllabification process.

These syllabified word fragments can be transliterated to Gujarati script using any approach and used to search matching documents.

E. g. Query word “Surat” should be syllabified as su-ra-t and it then it can be used for back transliteration to Gujarati script.

Rest of the paper is organized as follows: Section II describes related work. Section III describes our approach of transliteration. Section IV shows results and analysis and section V concludes the paper.

II. RELATED WORK

Shraddha Patel and Vaibhavi Desai in [1], in FIRE 2014 in Mixed Script IR task have employed combination of bi-gram and tri-gram with rule based approach for syllabification and transliteration. They have used Hindi as base language for transliteration to Gujarati language.

Substantial work is done in Indian languages for transliteration, especially in Hindi language. A Agarwal in [2] trained machine learning model for syllabification.

V Singhal et. al. in [3] developed rule based syllabification and rule based approach for English to Hindi transliteration. P Velunkar et. al. in [4] have also adapted rule based approach for syllabification. H Joshin et. al. in [5] used syllabification for transliteration of English to Hindi words.

For Gujarati language machine learning based approach is required for automatic Romanized word syllabification so that we can achieve more accuracy.

III. METHODOLOGY

In this section, we describe resources and implementation for syllabification of Romanized Gujarati words. For machine learning based transliteration of Roman to Gujarati script, syllabification is done using MaxEnt model and then word is transliterated. To accomplish this we have trained MaxEnt model.

a. Resources

For implementation of this methodology, 2169 manually syllabified Romanized Gujarati words are developed and used as resource to train MaxEnt model for machine learning based syllabification as shown in Table 1

da-sta-ve-j
sam-ban-dhi-t
sva-de-sh
pra-ni-yo
u-pa-yo-g
ja-gya
vya-va-stha-o
su-rya

Table 1 Sample Romanized Gujarati list of manually syllabified words used to train MaxEnt model for automated syllabification

b. Implementation

For transliteration from source script to target script, character by character mapping can be done. But one character in Source Roman script can’t be precisely mapped with one character of target script. E.g. Source Roman script character “a” can be mapped with “અ” or “એ” or “ા” in target Gujarati script. So we map syllabic components of word in source language to syllabic component of target language word. Syllabification refers to the process of phonetically decomposing the word. To syllabify a word, rule based approach and machine learning approaches can be implemented.

Machine learning based syllabification of Romanized Gujarati is done using trained MaxEnt model.

E.g. For transliteration, word ‘manushya’ is first syllabified using MaxEnt model as ma-nu-shya and then it is transliterated as મ-નુ-ષ્ય and back-transliterated syllabified fragments are combined back as મનુષ્ય.

We have trained MaxEnt model for automated word syllabification using manually syllabified 2169 Romanized Gujarati words list as shown in Table 1.

These transliterations may then be expanded with query in order to match retrieve documents written in Gujarati script. Figure-1 below shows flowchart of the CRF based machine transliteration for English to Gujarati script.

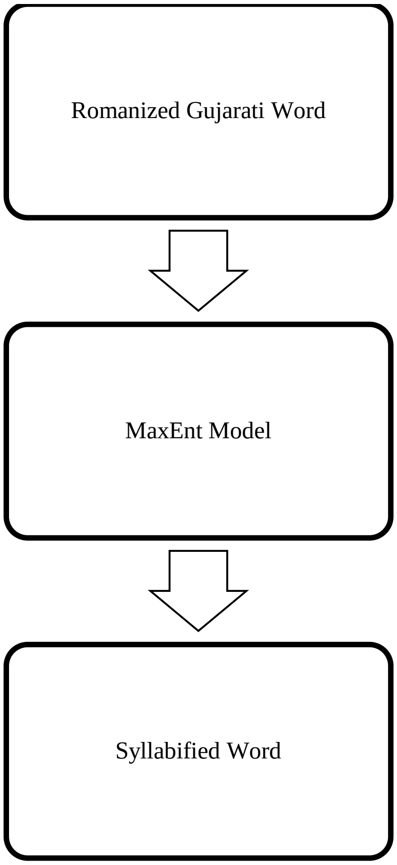

Fig.1: Flowchart of MaxEnt based syllabification of Romanized Gujarati words

IV. RESULTS AND ANALYSIS

In this section we show result comparison of syllabification with rule based approach and MaxEnt model trained by us.

We have conducted experiment to compare performance of rule based approach and MaxEnt model for Romanized words syllabification. Result of comparison for 456 Romanized Gujarati word syllabification is as given in Table 2 and figure 3.

Table 2: Result and comparison of Romanized Gujarati word syllabification methods

Method	Accuracy
Rule based	73%
Maximum Entropy	**87%**

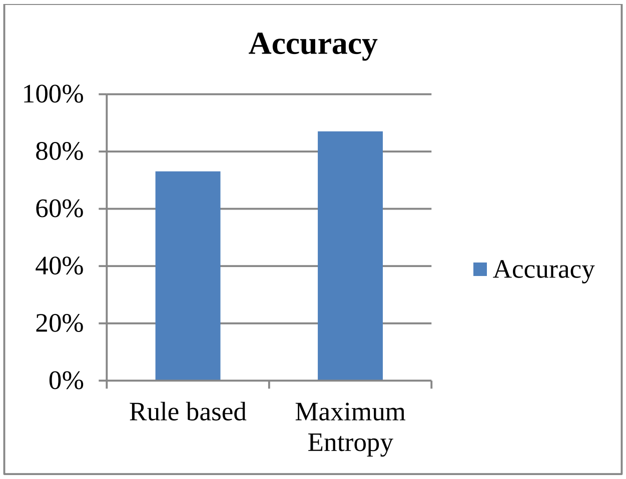

Fig.2: Comparison graph of Romanized Gujarati word syllabification methods.

As shown in result comparison of Table 2 and figure 2, using maximum entropy model we obtained 14% increase in accuracy for Romanized Gujarati word syllabification. As shown is result comparison graph in Figure-2, X-axis represents methodology applied for syllabification and Y-axis represents accuracy in percentage. It shows that using maximum entropy model we obtained 14% increase in accuracy for Romanized Gujarati word syllabification as compared to rule based approach.

These transliterations may be expanded with query in order to retrieve documents written in both Gujarati and Roman script. E. g. Query “mahatma ગાંધી” is expanded as “મહાત્મા + ગાંધી” and it retrieves documents containing words Mahatma and/or Gandhi in Gujarati language.

V. CONCLUSION

In this paper MaxEnt based machine learning approach is used for automatic syllabification of Roman script Gujarati word. Syllabified word fragments are then used for machine learning based transliteration. It will increase syllabification accuracy so that automatically accuracy of transliterated words will be increased.

This will make the scope of search broad for Gujarati language documents by giving user flexibility to search Gujarati documents using roman script to match mono-lingual Gujarati documents. In future we are going to enhance this work by adding more training data in order to cope up with wide vocabulary of both languages and also going to extend our work by adding more features.

VI. REFERENCES

[1]. Patel, S., and Desai, V. 2014. LIGA and Syllabification Approach for Language Identification and Back Transliteration. *Shared Task Reported by DAIICT in FIRE-2014.*

[2]. Agarwal, A. 2010. Transliteration involving English and Hindi languages using Syllabification Approach. *M.Tech thesis, Indian Institute of Technology, Bombay*. (Jan. 2010).

[3]. Singhal, V., and Tyagi, N. 2015. A Hybrid Approach of English – Hindi Named – Entity Transliteration. *International Journal of Advanced Technology in Engineering and Science*. 3, 2 (Feb. 2015), 580-587, ISSN: 2348 – 7550.

[4]. Verulkar, P., Balabantray, R. C., and Chakrapani, R. A. 2015. Transliterated Search on Hindi Lyrics. *International Journal of Computer Application*. 121, 1 (Jul. 2015), 32-37, ISSN: 0975 – 8887.

[5]. Joshi, H., Bhatt, A., Patel, H. 2013. Transliterated Search using Syllabification Approach. Forum for Information Retrieval and Evaluation. 2013