ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

PROCESSING OF ENGLISH-MALAYALAM CODE MIXED LANGUAGES USING NLP TECHNIQUES

Remya Sivan
E-Mail Id: remya.cse@sairamce.edu.in
Department of CSE, Sri Sairam College of Engineering, Bengaluru, India

Abstract- Code Mixing (CM) is an inclining research zone in Natural language processing (NLP) and it is characterized as blending of at least two dialects together in one record or sentence particularly in web-based life content like Facebook remarks, WhatsApp visits and so forth. In this day and age people are imparting their musings or insights through web-based life stages. In any case, rather than utilizing single language as their correspondence medium clients switch to and fro between various dialects, and this information is known as code mixed languages. Processing of code-mixed language is a challenging task in NLP. in this paper I attempted to learn about various assignments which are associated with preparing code mixed (CM) languages.
Keywords- NLP, Code mixed language, Word embeddings, Language identification, text normalization.

1. INTRODUCTION

Natural language processing is a territory of Artificial Intelligence (AI) and it is taken care of by Machine learning strategies. Machine is having the option to comprehend and process regular language information utilizing the information from man-made brainpower, etymology and software engineering is known as normal language processing. It has an assortment of utilizations like machine translation, document summarization, question answering, Information Retrieval and so forth. Since ordinary content information utilize single language as correspondence medium handling of it nearly simpler than internet-based life content which contains short messages, blending of more than one dialect, spelling blunders and so on. In this paper I studied handling of code – mixed language in web-based life content utilizing NLP. The internet based life content is unstructured and stirred up with more than one language together. These data are known as code mixed data. Essential assignment which engaged with handling of code exchanging are corpus creation, standardization(normalization), language identification, grammatical form labelling(tagging), parsing and so forth**.**

2. TYPES OF CODE-MIXED DATA

The various sorts of code-mixed information are.[1]

- Intra sentential
- Inter sentential
- Intra word level

Intra sentential- Inside one sentence more than one language known as intra sentential
Ex - Today (En) Njaan (Mal) coming (En).
English (En) and Malayalam (Mal) languages used within one sentence.
Inter sentential - Within one report various sentences are in various language known as inter sentential
Ex - Next Monday is too early (En). October avadhi kazhinjittu thudangiyal pore? (Mal)
Intra word level- Inside word more than one language utilized is known as intra word level exchanging Ex- Today (En) namukku (Mal) meetam “Meetam” is a word mix of English word “meet” and Malayalam suffix “aaam”.

3. PROCESSING OF CODE-MIXED DATA

Accessible apparatuses to process typical content fall flat with online networking content in light of its casual nature. It is possible that We need to adjust apparatuses which is as of now existing or make new tools to process code mixed data.[2][3]
There are two different ways to adjust NLP devices

- Perform standardization /Normalization
- Re-train the models inside the devices on annotated online networking content

Both have their own impediments so relies upon the application mix of both could be utilized.

3.1 Text Normalization -Standaization of Content

Content standardization is the procedure of change of casual language into formal language on which customary instruments were prepared. Word level Language distinguishing proof is the trailblazer of Text standardization. Before standardization label the words with relating language taggers. Content standardization makes two strides.

- Identify the orthographic mistakes
- Correct the mistakes

Accessible instruments to process typical content come up short with online life content due to its casual nature. Either adjust devices which is as of now existing or make new devices to process code blended data. There are two different ways to adjust NLP apparatuses

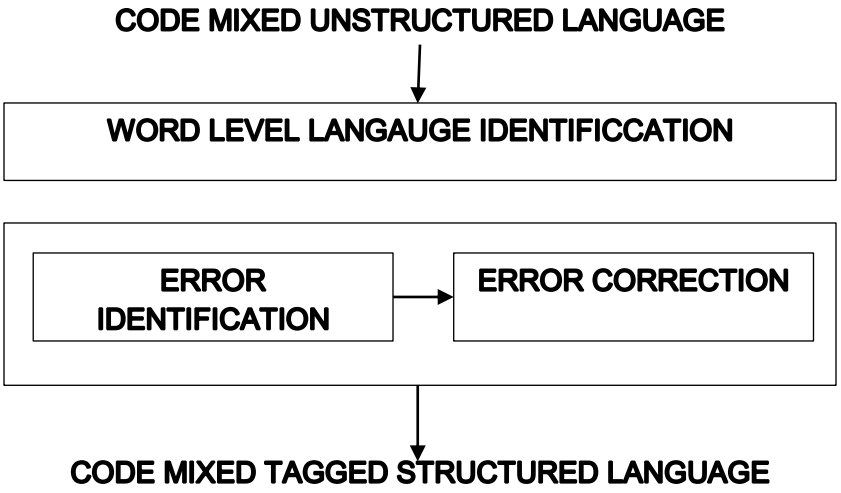

Fig. 3.1 Block Diagram of Standaization of Content

3.1.1 Word Level Language Identification

Language identification is a multiclass characterization issue. If there should be an occurrence of English Malayalam blended online life content, words are named "EN" for English "Mal" for Malayalam and "O" for Emoticons unique images and "UND" for mistakes.

Various methods for word level language recognizable proof are [4] character n-gram, Dictionary lookup, classification based on support vector machines, naive Bayes classifiers, sequence labelling with conditional random fields, logistic regression classifier. Aside from customary machine learning draws near, artificial neural networks models like recurrent neural system or convolutional neural network likewise utilized for word-based language identification [5]

Following are the challenges in word language identification

- Various transliterated spelling for given Malayalam word
  Ex: As per Cheran Unicode converter transliterated form നിങ്ങളുടെ is ningaLuTe. But people often use ningalude, ningalute etc
- Ambiguity:
  Language pairs share identical and similar words
  Ninte mole is looking good
  Ninte is tagged as Malayalam
  Mole can be tagged as either Malayalam or English
- Word is made from both English and Malayalam
  Ex: today namukku meetam
  Today is English
  Namukku is Malayalam
  Meetam is a single word which is combination of root English word “meet” and suffix Malayalam sound “aam”

Output of word language identification is tagging the word with corresponding language tagger and corresponding to Malayalam word Dravidian script and actual Unicode also written.

EX: From crowd arrow paranju

From EN crowd \En arrow\ MAL (aaro ആരോ ◌ാ) paranju \ MAL (paRanju പറഞ്ഞു)

3.1.2 Identify the Orthographic Errors

After word level language identification next step is to distinguish the orthographic mistakes in English words. Social media text incorporates, short messages or abbreviations(gm for good evening or gr8 for great)long messages or wordplay(good byeeeeeeee) spelling blunders or composing mistakes (remainder instead of reminder),

out of vocabulary words(innivation instead of innovation), wrong word in the context(son rather than sun), accentuation error(cant for can't), censor avoidance(f***),and emoticons😊().Identify and classify all these unpredictable words in English language .The ordinary methodology for content standardization is make dictionary of right words and distinguish in vocabulary and out vocabulary as far as word reference.

The fundamental Dictionary lookup methods are Finite automata, hashing and binary search tree. The subsequent strategy is n-gram method. N-gram works dependent on the probabilities. The model which appoint probabilities to a set of characters or succession of word is known as language model .N-gram is one sort of language model .N-gram is a lot of sequential characters from the given content with estimation of N .If N=1 then it is unigram and N=2 and 3 it is bigram and trigram separately. Likelihood is determined by the condition P(w/h), probability of a word w given history. History can be unigram bigram or trigram.

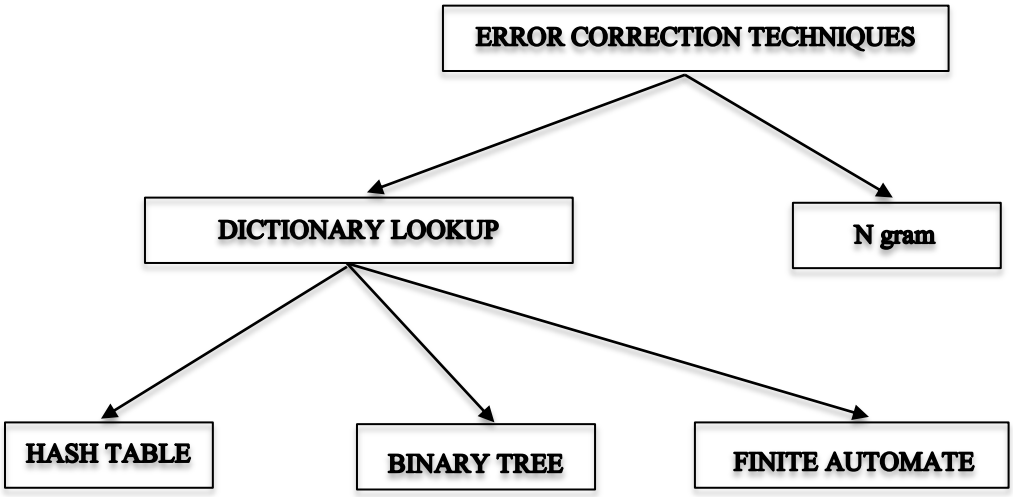

Fig. 3.2 Error Identification Mehtods

3.1.3 Error Correction

There are numerous mistake adjustment methods. The fundamental methods are [7]

- Minimum edit distance
- Similarity keys
- Rule based
- Probabilistic techniques
- N-gram
- Neural Network

Utilize any of these strategies to address the mistakes. Subsequent to rectifying the mistakes the code blended labeled organized language can utilize existing NLP apparatuses to process.

3.2 Re-Train the Models inside the Tools via Web-based Networking Media Content

The accessible devices of NLP are prepared on organized cautiously altered information. These instruments won't give legitimate outcome on unstructured code-mixed language. The one technique to utilize the as of now accessible NLP devices are retrain the models utilizing social media text. In the event that the preparation tests are accessible, at that point re-preparing NLP devices are simple else it is tedious job [2].

CONCLUSION

Right now, I have learned the strategies to process the codemixed English Malayalam dialects and distinguished the difficulties in English Malayalam code mixed language. In future I will actualize the calculation dependent on recurrent neural network to recognize word-based language distinguishing proof and standardization of code mixed Manglish language.

REFERENCES

[1] Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text Amitava Das Björn Gambäck University of North Texas Norwegian University of Science and Technology Denton, Texas, USA Trondheim, Norway.

[2] NLP for Social Media Lecture 2: Text Normalization Monojit Choudhury Microsoft Research Lab.
[3] Natural Language Processing for Social Media, Second Edition Atefeh Farzindar and Diana Inkpen 2017.
[4] Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text Amitava Das Björn Gambäck University of North Texas Norwegian University of Science and Technology Denton, Texas, USA Trondheim, Norway.
[5] Neural Network Methods for Natural Language Processing Yoav Goldberg Synthesis Lectures on Human Language Technologies, April 2017, Vol. 10, No. 1, Pages 1-309.
[6] Code Mixing: A Challenge for Language Identification in the Language of Social Media Utsab Barman, Amitava Das† , Joachim Wagner and Jennifer Foster CNGL Centre for Global Intelligent Content, National Centre for Language Technology School of Computing, Dublin City University, Dublin, Ireland †Department of Computer Science and Engineering University of North Texas, Denton, Texas, USA.
[7] Spell Checking Techniques in NLP: A Survey ,Neha Gupta Pratistha Mathur Department of Computer Science, Banasthali Vidyapith, Jaipur, Rajasthan, India.