Context Specific Spell Checker for Hindi

Group 9

Objectives

- Act as a spell checker
  - Being able to detect misspelled words
  - Provide suggestions for the replacement of the misspelled word
- Find phonetically similar words
  - Produce words that are similar sounding to a word
  - Rank these words on the basis of context

Types of Errors

Spelling Mistakes

- चुनौतयों - चुनौतियों
- सीरफ – सिरफ
- बराने – बढ़ाने
- आतक – आतंक

Phonetically Similar Words

- संग – संघ
- भजन – वजन
- बीस – विष
- पक्का – पका

Intention was to generate mistakes that are commonly made, as it would help in validating the model.

Implementation

Objective :

- Keep it simple
- Not complicate things
- Make a model with several adjustable parameters

Major Checkpoints :

- Bucketing phonetically similar words together
- Performing unicode correction and generating edit distance
- Finding the the contextual score on the basis of bigram frequencies
- Generating test data for checking correctness

Implementation

- Phonetic Mapping
  - Character Classification :
    - Mapping each character of the hindi alphabet to a particular code
    - Similar sounding groups get the same code
  - Bucketing
    - On the basis of the code generate buckets of similar sounding words

- Unicode Correction
  - Sentences into list of words (= tokens)
  - Broke a word into list of character
  - Consonant + Matra -> Consonant + Halant + Matra => की -> [क + 'ि' , ई]
  - Consonant -> Consonant + Halant + 'अ' => क -> [क + '्' , 'अ']

Implementation

- Edit Distance :
  - Used Levenshtein Distance
  - Tunable cost parameters :
    - Addition, Deletion, Swapping
  - Return large distance when the difference in length of words is greater than 5 (average length of word)

- Contextual Scoring
  - Generated Bigram Frequencies
  - Scoring Metric for a tuple(w1, w2) : (w_i, w_j)

$$score(w_i, w_j) = \frac{bi\_freq(w_i, w_j)}{\sqrt{freq(w_i) * freq(w_j}}$$

Testing

- We created two data sets :
  - Sentences were curated inhouse by looking up most frequent mistakes
  - Spelling Mistake List
    - इसीलिए –इसलिए लाल रक्त कोशिकाओं की संख्या बराने – बढ़ाने के लिए आयरन वाले फूड्स खाएं
    - उनका सेनापति सहीद – शहीद होगया
  - Phonetic Mistakes List
    - शरीर के भजन – वजन का 10 प्रतिशत वार – भार सिर्फ खून का होता है
    - सांप का बीस – विष घातक होता है

Testing

Phonetic Mistakes :

सांप का बीस - विष	बीन, बीच, बीज
आम पक्का - पका	ढक्का, धक्का
राष्ट्रीय सेवक संग - संघ	संघ, संध, संत
चुनौतियों के हाथ - साथ	साथ, हार
मनाया खायेगा - जायेगा	जायेगा, आयेगा
शरीर के भजन - वजन	वजन, भवन

- Out of 18 mistakes made, 14 were correctly identified
- Comparatively cases where the phonetically similar words had considerable spelling changes did not show the correct replacement
  - However the words recommended showed considerable context relevance and similarity to the original word used
- In some cases the correct words was provided though with lower ranking

Testing

Spelling Mistakes :

सेब अध काटके - आधा	अब, अंत, अर्ध
समसयाओं - समस्याओं	सरगनाओं, समकक्षों
एक मजदार रोड - मजेदार	मजेदार, मजदूर
चुनौतयों के साथ - चुनौतियों	चुनौतियों, चुनौतीयों
संख्या बराने के - बढ़ाने	कराने, बढ़ाने
कार्यकम आयोजित कया	कहा, किया

- Out of 16 mistakes made, 14 were correctly identified
- The correct words were present in the top 5 ranked words, however they were not the first preference in many cases
- The words that were correctly present, also had a list of word they could be replaced by.
    - The correctly placed word ranked lower than other words
- Spelling mistakes that produced a high edit distance were ignored, and hence produced inaccurate results

Further Improvements

- Improving contextual performance
  - Try with different n-gram frequencies
- Tweaking Edit Distance by using Similarity Matrix
  - Based on similarity in characters, tweak the swapping penalty
- Using Edit Distance and Contextual Score to develop a combined metric
- Testing Metric
  - Current test indicate proof of concept and show correct working
  - Generate test set to determine accuracy, precision etc.
- On the go suggestions
  - Currently run iteratively on every word of the sentence
  - Could implement an on the go suggester

References

Below are the papers we referenced :

1. “UTTAM”: An Efficient Spelling Correction System for Hindi Language Based on Supervised Learning https://dl.acm.org/doi/10.1145/3264620
2. Hindi Spell Checker, https://cse.iitk.ac.in/users/cs365/2013/submissions/~pulkitj/cs365/project/report.pdf
3. Design and Implementation of HINSPELL -Hindi Spell Checker using Hybrid approach. https://ijsrm.in/index.php/ijsrm/article/view/102
4. A study of spell checking techniques for Indian Languages http://jkhighereducation.nic.in/jkrjmcs/issue1/15.pdf