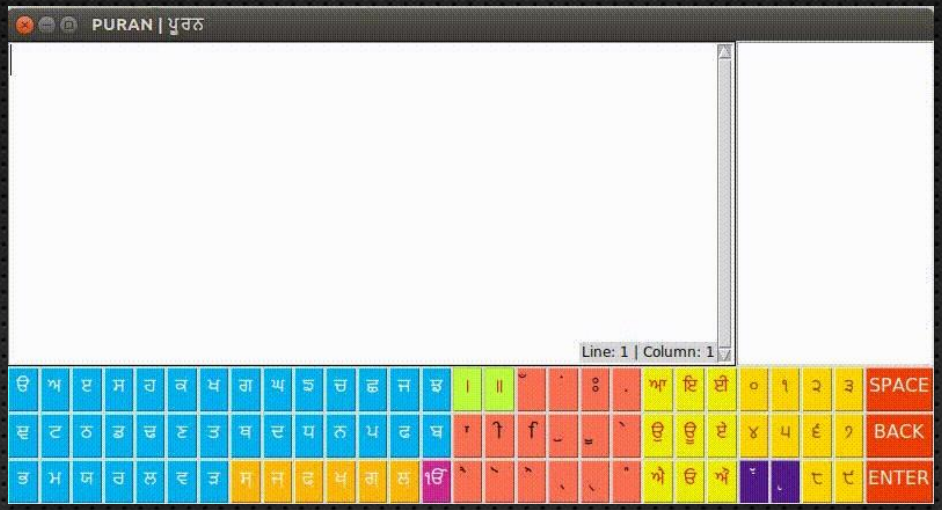


PURAN

WORD PREDICTION SYSTEM FOR PUNJABI LANGUAGE NEWS

Gurjot Singh Mahi, Amandeep Verma
Department of Computer Science, Punjabi University, Patiala, Punjab, India

19 JANUARY 2019
ICDMAI-2019, Kuala Lumpur, Malaysia

Agenda of the Presentation

This paper presents an outline of the PURAN: A state-of-the-art word prediction system for Punjabi language news.

This presentation elaborates the word prediction system architecture in detail.

Also, demonstrates that the PURAN has achieved highest Hit ratio in Regional news genre followed by National news genre by making lowest average keystrokes in the said categories of news.

Contents

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am l

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living in L

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living in London

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living in London

Word Prediction System

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words.

Hi I am living in London

Several software applications like PROFET, FASTY, PAL were proposed to predict the word in English and other European languages.

Large character database, different vowels symbols and complex language syntax makes text composition difficult in Indian context.

This study intends to present a word prediction system - PURAN, designed to predict the words in News items for Punjabi language.

[Carlberger, J. Carlberger, T. Magnuson, M. S. Hunnicutt, S. E. Palazuelos-cagigas, and S. A. Navarro, "Profet, A New Generation of Word Prediction: An Evaluation Study," Nat. Lang. Process. Commun. aids, pp. 23–28, 1997.]
[J. Matiasek, M. Baroni, and H. Trost, "FASTY — A Multi-lingual Approach to Text Prediction," in International Conference on Computers for Handicapped Persons, Berlin, Heidelberg: Springer, 2002, pp. 243–250.]
[A. Newell, J. Arnott, L. Booth, W. Beattie, B. Brophy, and I. Ricketts, "Effect of the 'PAL' word prediction system on the quality and quantity of text generation," Augment. Altern. Commun., vol. 8, no. 4, pp. 304–311, Jan. 1992.]

Background

The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

Methodology

Statistical Predictor

The statistical predictors rely on the n-gram language model for language information in statistical word prediction system

Syntactic Predictor

Rule based syntactic structures of a natural language are followed for the design of syntactic word prediction systems.

Background

The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

Methodology

Statistical Predictor

The statistical predictors rely on the n-gram language model for language information in statistical word prediction system

Word	Frequency
the	746240010
you	131164406
there	23199253
may	16406340
will	51209514

Syntactic Predictor

Rule based syntactic structures of a natural language are followed for the design of syntactic word prediction systems.

PRP$/ My NN/ Name VBZ/ is NNP/ Gurjot ./ .

Background

The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

Methodology

Statistical Predictor

The statistical predictors rely on the n-gram language model for language information in statistical word prediction system

Word	Frequency
the	746240010
you	131164406
there	23199253
may	16406340
will	51209514

Word Frequency is taken into account

Syntactic Predictor

Rule based syntactic structures of a natural language are followed for the design of syntactic word prediction systems.

PRP$/ My NN/ Name VBZ/ is NNP/ Gurjot ./ .

Part of Speech is taken into account with frequency

Background

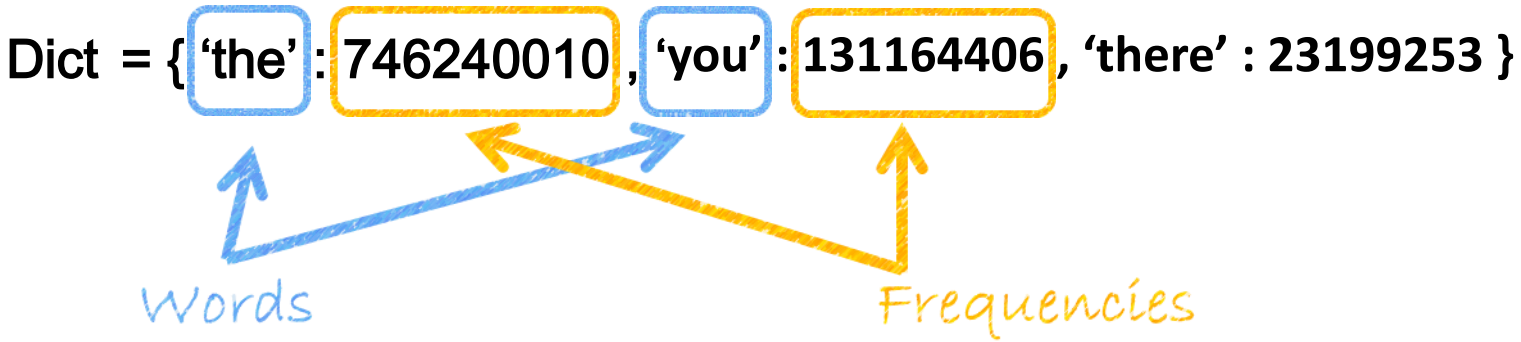
The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

Dictionary

An arrangement of the words with their frequencies or probability values requires a data structure for enhanced and firm results.

Dictionary is a general concept that is used to manage the **keys** and its **values** in an efficient manner.



Concept of Dictionary in Python Language

Background

The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

User Interface

The user interface provides an interactive space for a list of suggestions to appear from which user selects the most appropriate word that matches the user specified prefix.

Prediction Proposals				
pay	pain	pack	page	paint

[N. Garay-Vitoria and J. Abascal, "Text prediction systems: A survey," Univers. Ac-cess Inf. Soc., vol. 4, no. 3, pp. 188–203, 2006.]

Prediction proposals
pack
page
pain
paint
pay

1st	2nd	4th	B	N	U	B	Z
3rd	5th	A	R	T	Y	K	
blan	O	L	C	H	X		
S	D	M	V	W			
I	P	G	.				
Q	J	,					
F	:						
;							

Point-of-gaze plays an important role in increasing the production rate of a word predictor.

Minimum eye and hand movement should be supported by the user interface

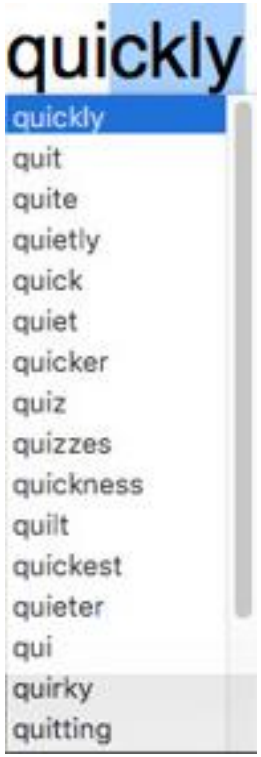
[A. Newell, J. Arnott, L. Booth, W. Beattie, B. Brophy, and I. Ricketts, "Effect of the 'PAL' word prediction system on the quality and quantity of text generation," Augment. Altern. Commun., vol. 8, no. 4, pp. 304–311, Jan. 1992.]
[N. Garay-Vitoria and J. Abascal, "User interface factors related to word- prediction systems," in Proceedings of the 7th International Conference on Work With Computing Systems WWCS2004., 2004, pp. 77–82.]
[M. K. Sharma and D. Samanta, "Word Prediction System for Text Entry in Hindi," ACM Trans. Asian Lang. Inf. Process., vol. 13, no. 2, pp. 1–29, 2014.]

Background

The idea of the development of a word prediction system was started way back in 1968 when (H. C. Longuet-Higgins and A. Ortony, 1968) published the technique to reduce the keystrokes for completing a word.

Four main factors affecting word prediction are

Number of suggestions



With the use of more suggestions in the suggestion list, a predictor can avail higher hit ratio, but will also increase the cognitive load on the user

(N. Garay, J. Abascal, and L. Gardeazabal, 2002) points that suggestion list with ten suggestions provide stability between keystrokes and cognitive cost.

[N. Garay-Vitoria and J. Abascal, "Text prediction systems: A survey," Univers. Access Inf. Soc., vol. 4, no. 3, pp. 188–203, 2006.]
[N. Garay, J. Abascal, and L. Gardeazabal, "Evaluation of prediction methods applied to an inflected language.," in International Conference on Text, Speech and Dia-logue, 2002, pp. 397–403.]

System Architecture

To perform the prediction of the word - ω, the system architecture is distributed into two phases,

1) Corpus creation and statistical inference — Phase 1

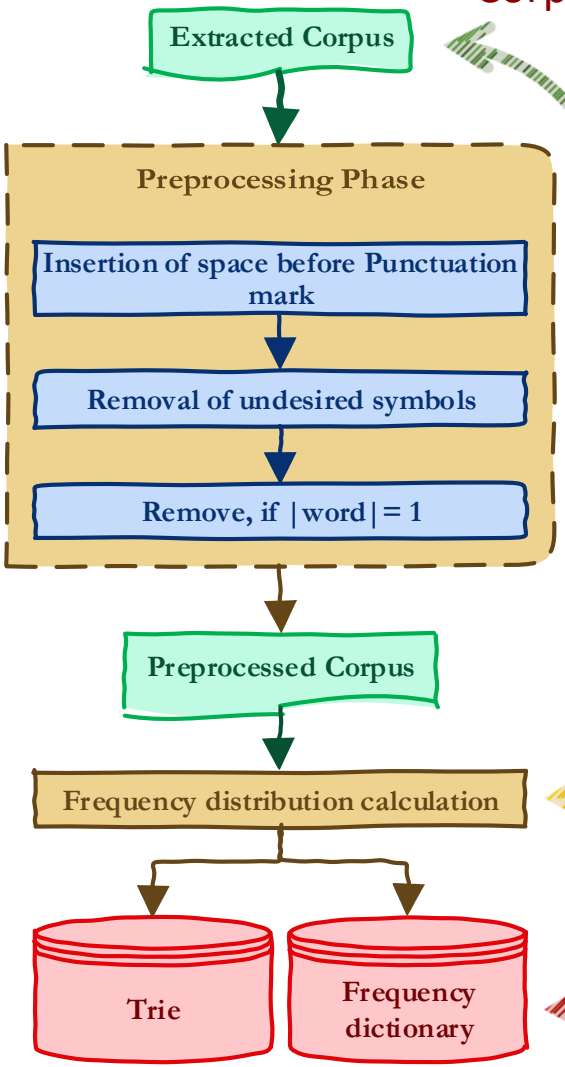
2) Word prediction — Phase 2

The phase 1 of corpus creation and statistical calculation is performed only once.

Phase 2 is initiated whenever the user enters any new prefix or update the prefix with a new set of characters.

System Architecture

Corpus creation and statistical inference



The initial phase of the system architecture is divided into four diverse steps,

1) Extraction of Punjabi news corpus

2) Preprocessing of the corpus

3) Calculation of frequency distribution

4) Formation of words Trie and frequency dictionary.

Fig. 1. Corpus creation and statistical inference

System Architecture

Corpus creation and statistical inference

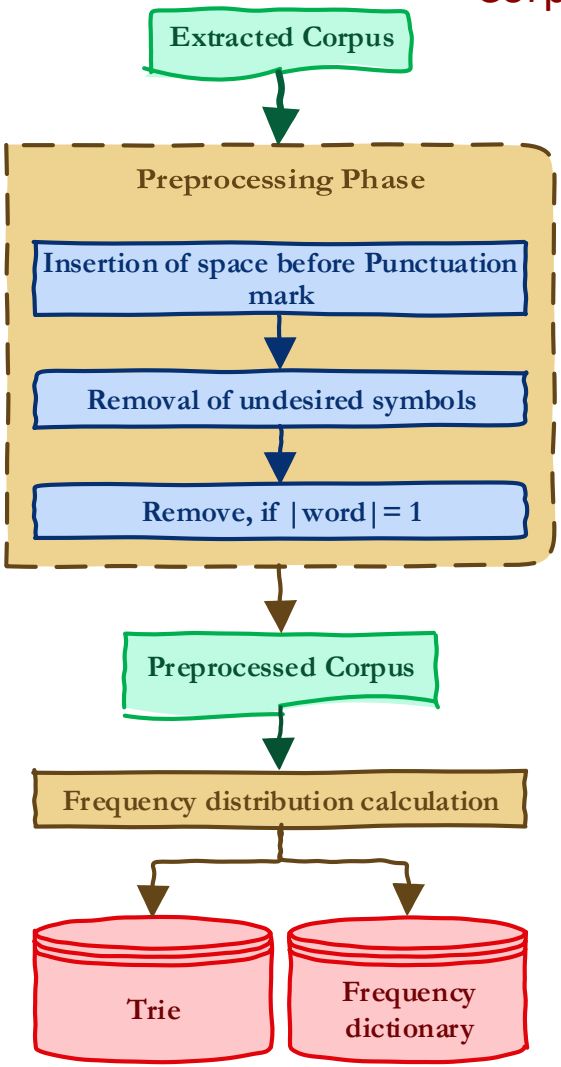
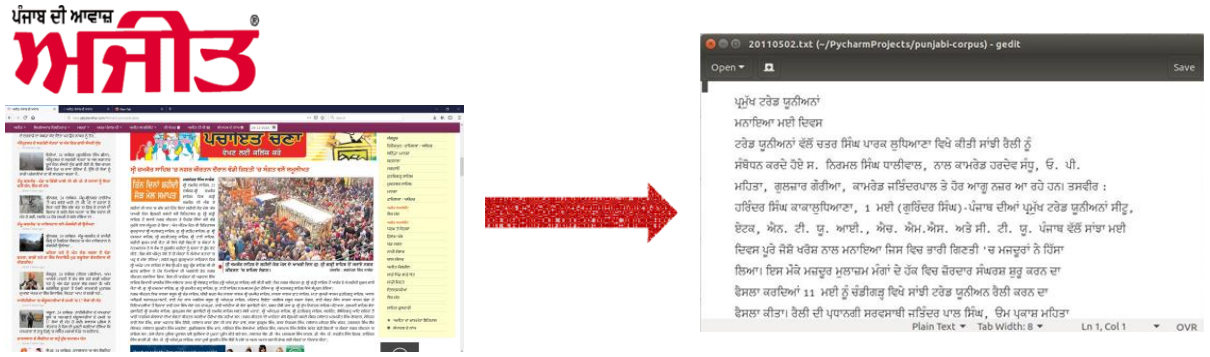


Fig. 1. Corpus creation and statistical inference

1) Extraction of Punjabi news corpus



http://beta.ajitjalandhar.com

News corpus is extracted from Ajit Newspaper website. The extracted corpus contains 5 genres of news i.e. Business, International, National, Regional, and Sports.

2) Preprocessing of the corpus

Extracted corpus is pre-processed before any statistical inference could be obtained.

Statistics	Before Preprocessing	After Preprocessing
Total words	1408041	1357443
Total characters	8795714	7075784
Average length of words	4.252	4.212
Unique words	65551	48521
Mode of words	4	4

Table 1. Summary of extracted corpus statistics

Extracted Corpus — Processing function

$$Preprocess\,(c) = \acute{c}$$

Processed Corpus With N variables

System Architecture

Corpus creation and statistical inference

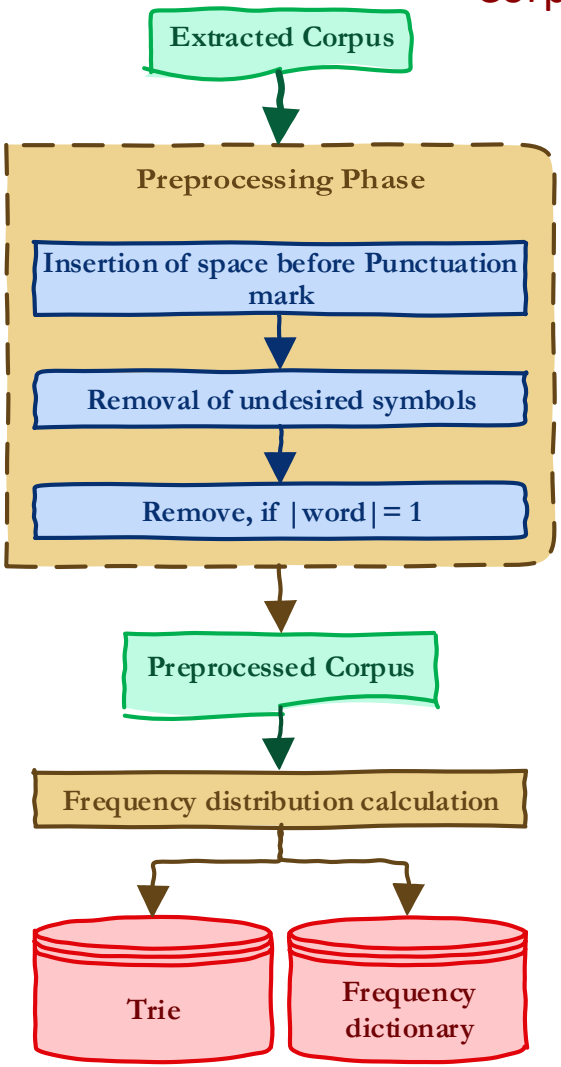


Fig. 1. Corpus creation and statistical inference

3) Calculating frequency distribution

Another significant step in the system architecture is to make the statistical inference.

Frequency distribution is calculated from the preprocessed corpus ($\acute{c}$).

Formally it can be stated as,

set of discrete variables

$$VAR_{\acute{c}} = \{x_0, x_1, x_2, \dots, x_n\}$$

n denotes the number of discrete variables x in processed corpus **ć**.

The frequency of i^{th} discrete variable in the corpus $\acute{c}$ can be given by,

$$f_0 + f_1 + f_2 + \cdots + f_n = N$$

Frequency set

$$FREQ_{\acute{c}} = \{f_0, f_1, f_2, \dots, f_n\}$$

Where, f_i is the frequency for i^{th} discrete variable for $i = 0,1,2, \dots, n$ in $VAR_{\acute{c}}$.

System Architecture

Corpus creation and statistical inference

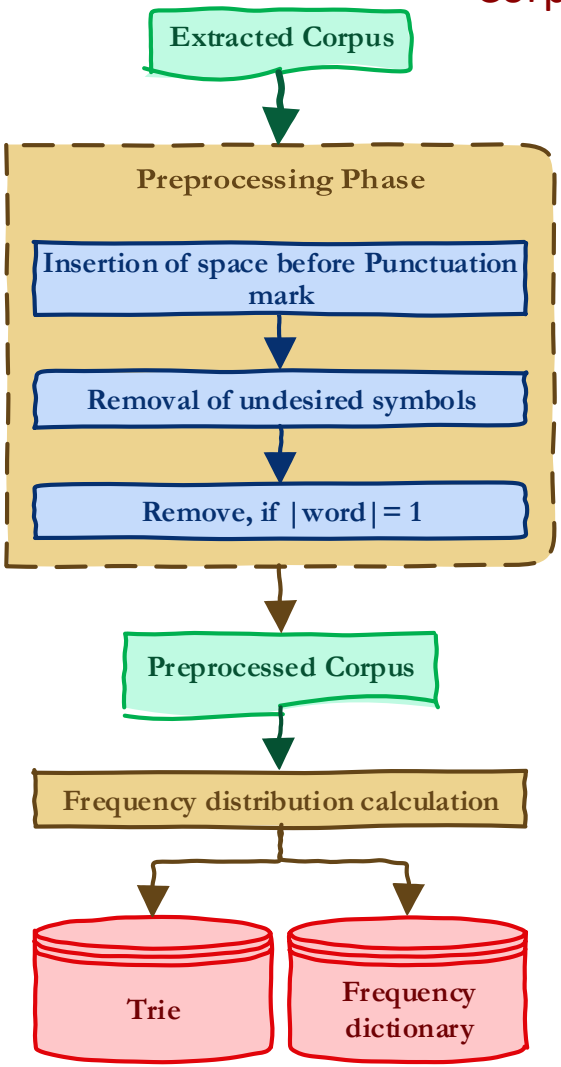
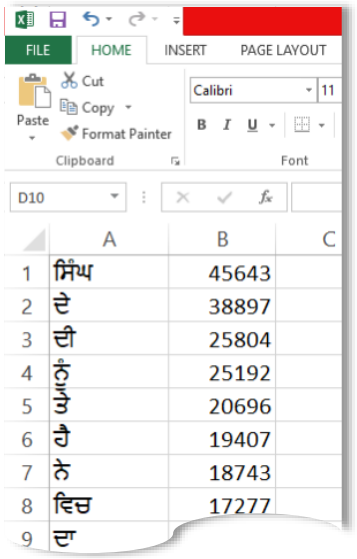


Fig. 1. Corpus creation and statistical inference

4) Formation of words Trie and frequency dictionary

Set $VAR_{ć}$ and $FREQ_{ć}$ are utilized in the final step for the formation of word Trie and frequency dictionary.



Punjabi News words frequency excel file

System Architecture

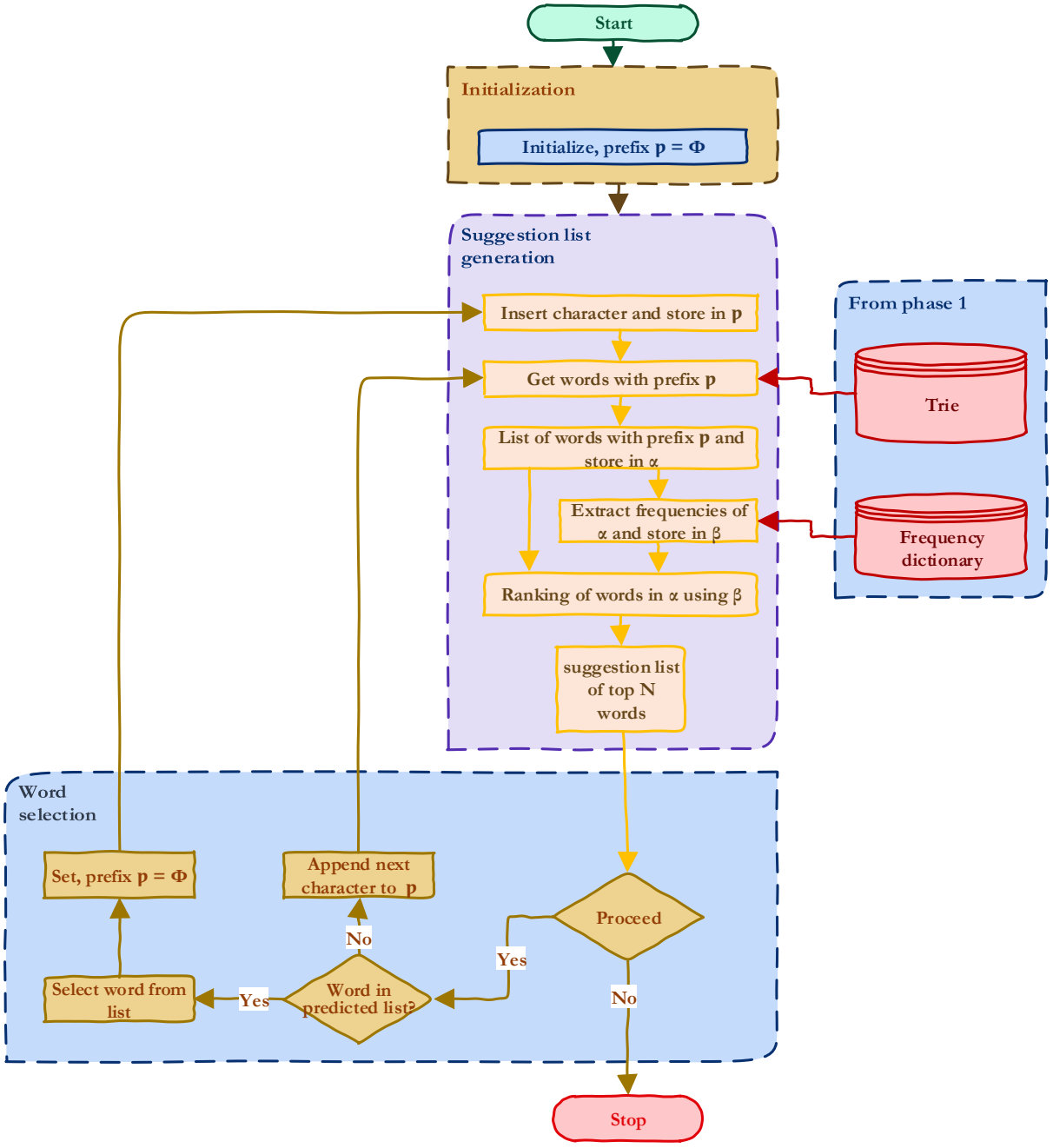
Word prediction

1. Initialization of prefix

The initial phase in this step starts by initializing the prefix with null.

If p denotes the user entered prefix, then this step can be stated as,

$$p = \Phi$$



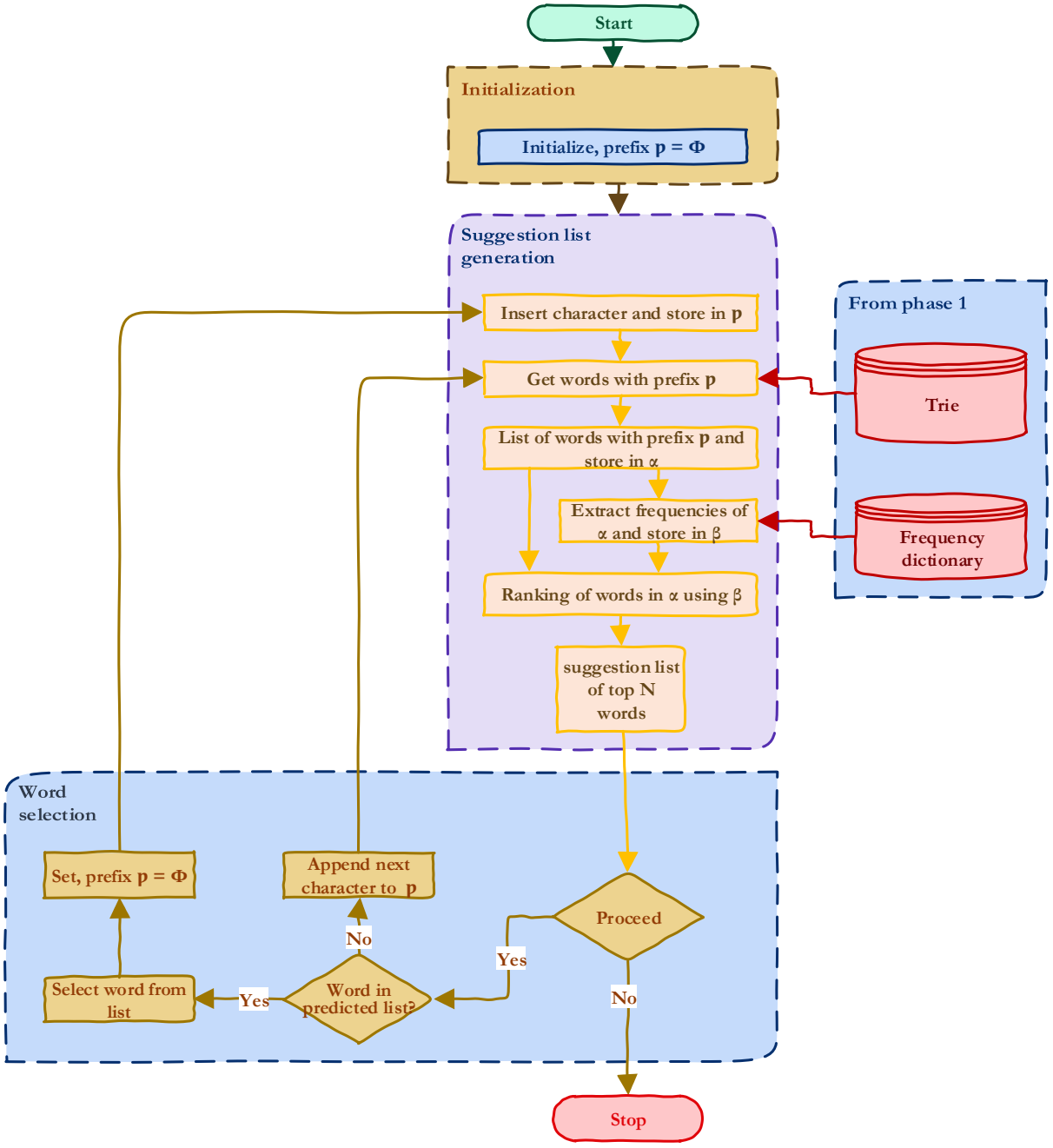
System Architecture

Word prediction

2. Suggestion list generation

User intended first character is stored in prefix – p and is used to extract all the similar words in list– α, from $TRIE_{PUN}$, which more formally is known as set $VAR_{\acute{c}}$.

Furthermore, list of words in α are used to extract frequencies in another list– β, using frequency dictionary $DICT_{PUN}$,which more formally is stated as set $FREQ_{\acute{c}}$



System Architecture

Word prediction

2. Suggestion list generation

Let $L_{PRED}(p)$ denotes the set of extracted words from $TRIE_{PUN}$ such that,

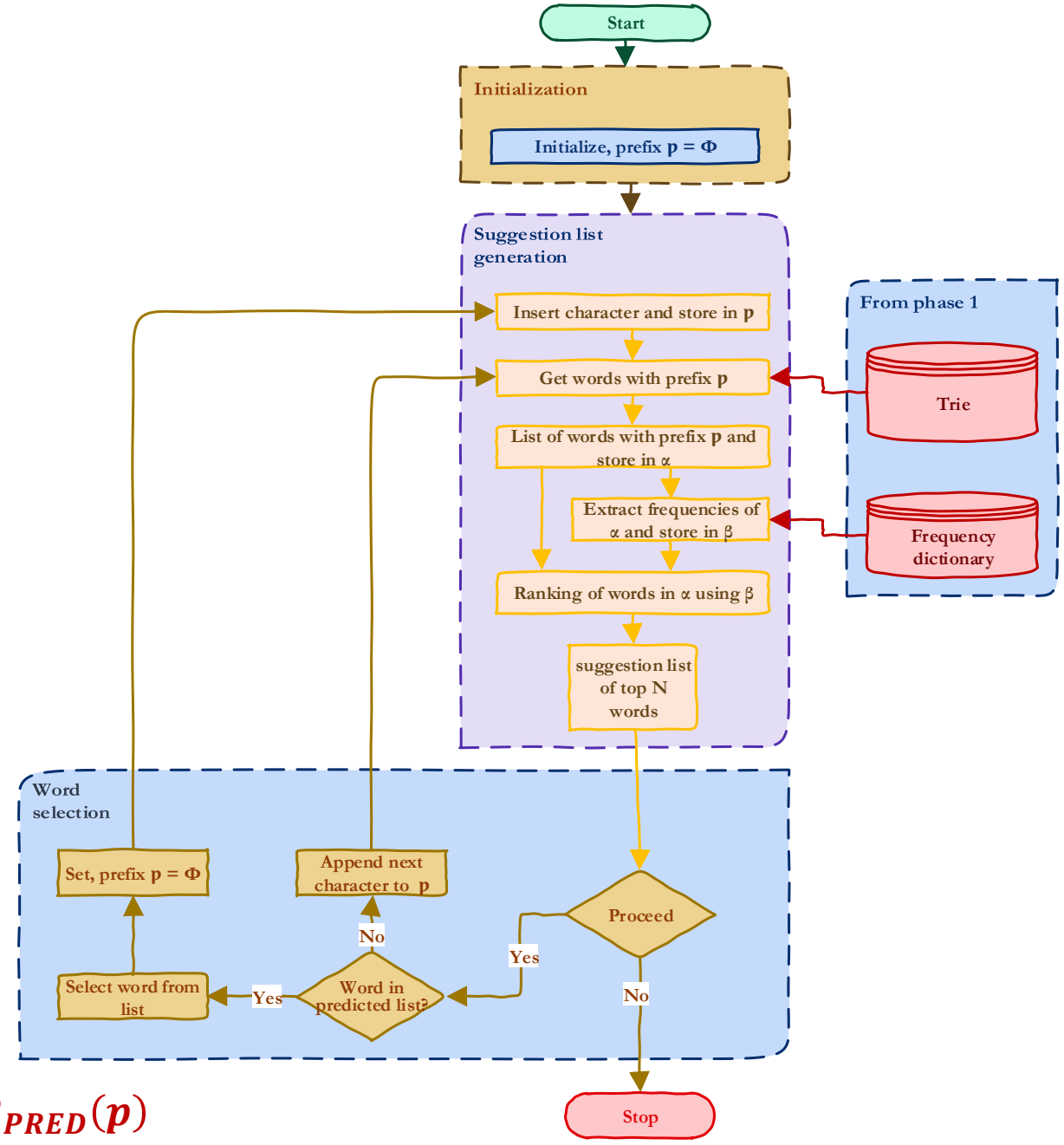
$$L_{PRED}(p) \in VAR_{ć}$$

set of extracted words from $\boldsymbol{TRIE_{PUN}}$

Total cost of $\boldsymbol{R_{PRED}(p)}$

$$COST(R_{PRED}(p)) = \sum_{\omega \in R_{PRED}(p) \subset L_{PRED}(p)} COST(\omega)$$

cost of suggesting $\boldsymbol{\omega}$ *in* $\boldsymbol{R_{PRED}(p)}$



System Architecture

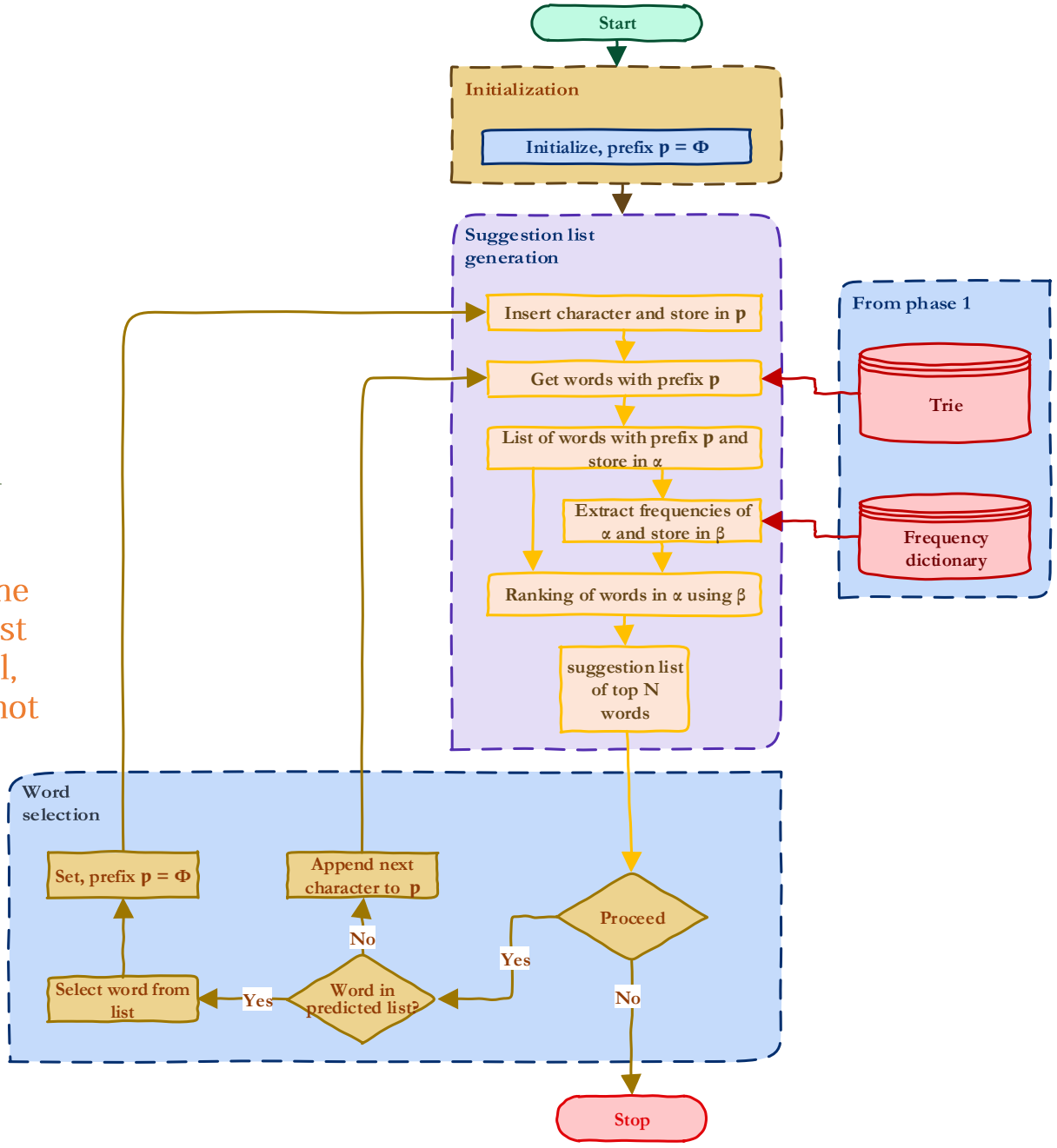
Word prediction

3. Word selection

The last step is to select the user intended word – ω from N words in the list α.

If the intended word – ω is found the list, the word is selected from the list and prefix is again initialized to null, i.e. p = Φ. If user intended word is not found in the list, next character is appended to prefix – p.

Step 2 and 3 of phase 2 is iterated till user intended word – ω is not found for the updated value of p or till system does not exit.



Performance metrics

Hit ratio is used to describe the reliability of the word prediction system. Higher the hit ratio, higher will be the credibility of the prediction system to predict the correct word.

$$Hit\ ratio = \frac{number\ of\ times\ words\ is\ predicted}{total\ number\ of\ written\ words}$$

Keystroke saving is used to measure the actual saving of keystrokes. It measure the number of keystrokes saved.

$$Keystroke\ saving = 1 - \frac{number\ of\ keystrokes\ made}{total\ length\ of\ word}$$

Average rank calculation mechanism is utilized to compute the average rank of the predicted words in the vertical list of suggestions.

$$Average\ rank = \frac{ranks\ total}{number\ of\ words}$$

Average keystroke calculation mechanism is utilized to recognize the average number of keystrokes entered for each word during the prediction testing.

$$Average\ keystroke = \frac{total\ keystrokes\ made}{number\ of\ words}$$

Testing dataset and system configuration

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import re
import xlsxwriter
import urllib.request
from bs4 import BeautifulSoup
from urllib.parse import quote
from urllib.request import urlopen

'''
initialize file name for news data management
'''
FactSheet = "C:\\Users\\STANZIN DAWA\\PycharmProjects\\web_scrapper\\data_files\\
workbook = xlsxwriter.Workbook(FactSheet)
worksheet1 = workbook.add_worksheet()
worksheet1.write(0, 0, "Text_File_No")
worksheet1.write(0, 1, "Title")
worksheet1.write(0, 2, "Genre")
worksheet1.write(0, 3, "Time")
worksheet1.write(0, 4, "Unique Words")


'''initialize jagbani newspaper URL'''
html = urlopen("http://jagbani.punjabkesari.in/latest.aspx")
jagbani_pages= []

def initialize():
    if html is None:
        print("URL is not found")
    else:
        '''
        This parts extracts all the link for in jagbani website pages
        '''
        bsObj = BeautifulSoup(html, "html.parser")
        front_url = "http://jagbani.punjabkesari.in"
        div = bsObj.findAll('div', attrs={'class': 'kjpage'})
        for page_no in div:
            links = page_no.find_all('a')
            for link in links:
                jagbani_pages.append(front_url+link['href'])
```

Jagbani Website Crawler
(https://github.com/GurjotSinghMahi/jagbani_website_crawler)

Jagbani New Corpus

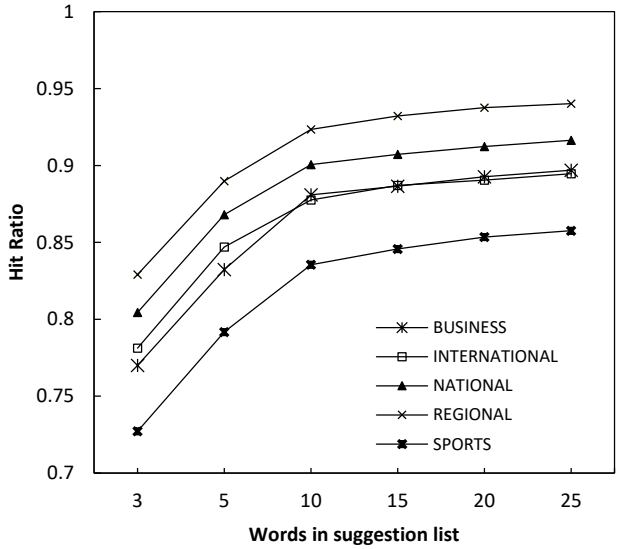
Genre	News items
Business	22
International	65
National	48
Regional	115
Sports	31

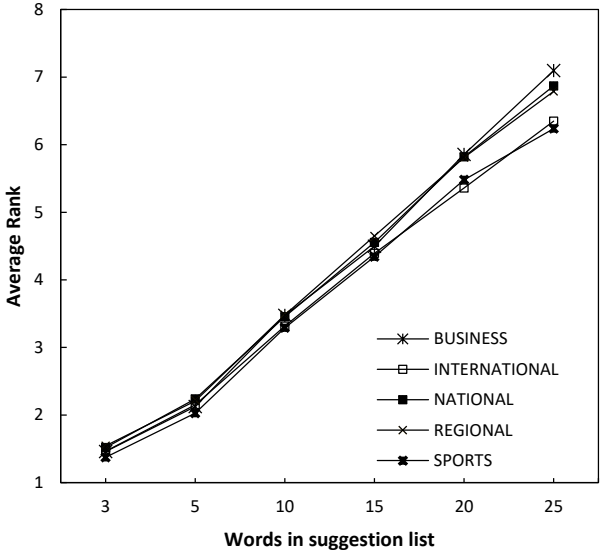
Table 3. Classification of news in testing corpus

The system architecture was implemented on the Windows 10 Operating system, Intel Core i5-6200U CPU 2.40 GHz with 8 GB RAM. The system is designed using Python programming language. NLTK and URLLIB packages are used for the implementation of said system and design of Web crawlers.

Results and discussion

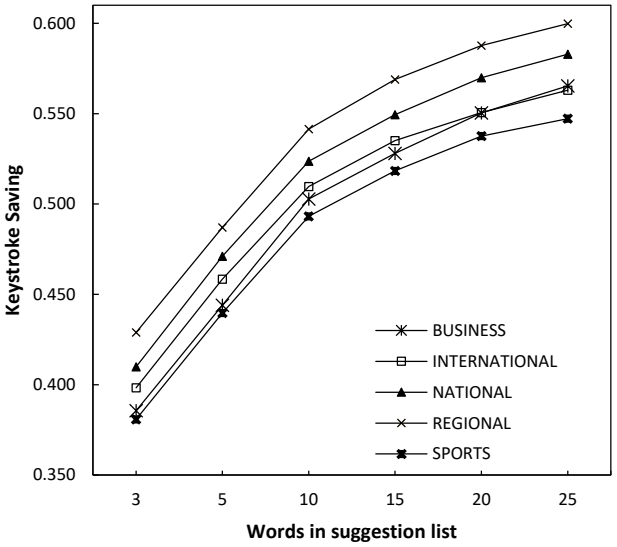
(a) Hit ratio for different Genres of news

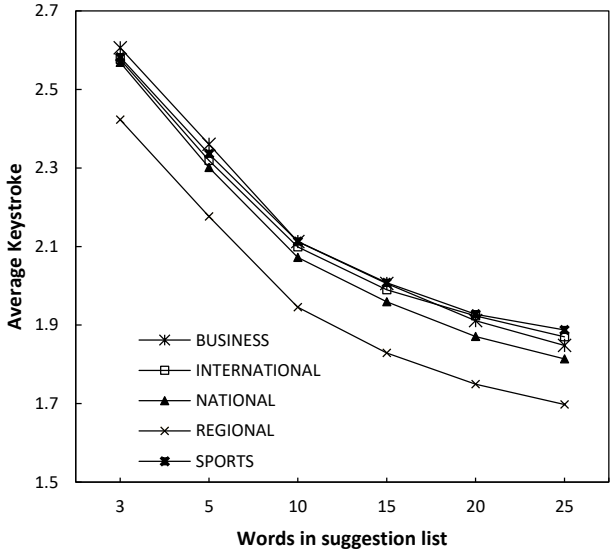




(b) Average Rank for different genres of news

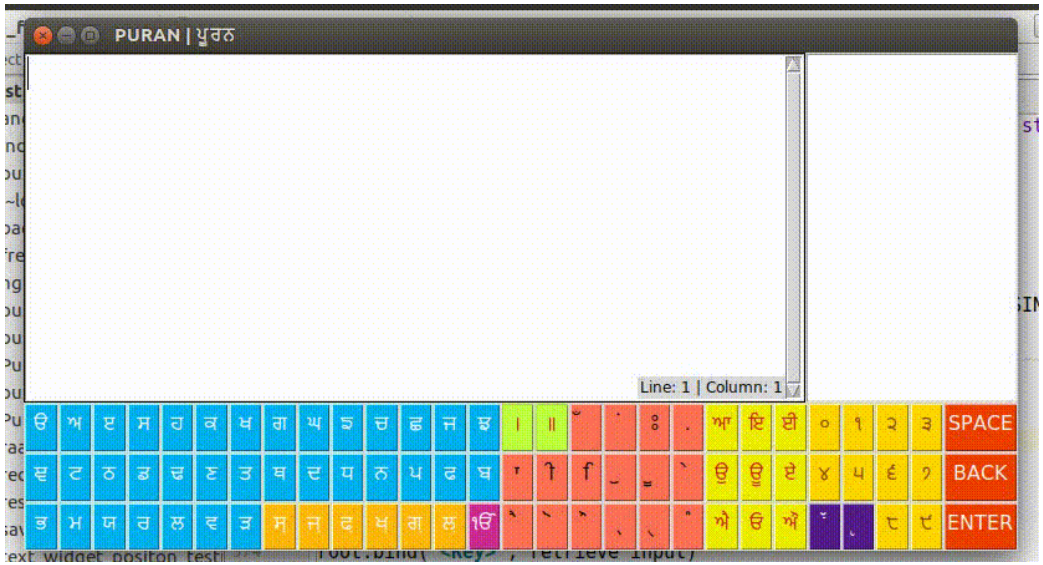
(c) Keystroke saving for different genres of news





(d) Average Keystrokes for different genres of news

Conclusion



1. **The overall applied architecture used for suggesting the user the most appropriate word based on the prefix provided by the user was demonstrated.**
2. **The system performance is tested on the various benchmark metrics like *Keystroke saving*, *Hit ratio*, *Average rank* and *Average keystrokes* for rigorous review of the proposed system to examine its credibility in Punjabi news category.**
3. **PURAN prediction system works well in the categories of Regional and National news genres followed by International, Business and Sports genres.**
4. **The system has achieved 88.38% Average Hit ratio with 51.42% Average keystroke saving for N=10.**

Thank you!

gurjotmahi28@gmail.com
vaman71@gmail.com

The slides will be posted on:
https://github.com/GurjotSinghMahi/

The idea of presentation was taken from
https://piotrmirowski.files.wordpress.com/2016/11/piotrmirowski_2016_meetup.pdf

Punjabi IPA conversion code is available at:
https://github.com/GurjotSinghMahi/PUNJABI-IPA-CONVERTER