ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(27), DOI:10.17485/ijst/2015/v8i27/81667, October 2015* ISSN (Online) : 0974-5645

Punjabi Dialects Conversion System for Malwai and Doabi Dialects

Arvinder Singh and Parminder Singh*

Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana - 141006, Punjab, India; arvinder006@gmail.com, parminder2u@gmail.com

Abstract

In recent times the rate of colloquial and informal Punjabi text has increased dramatically. People are using the web as medium for expressing their ideas, thoughts and emotions, usually in form of colloquial articles and blogs. Punjabi has a wealth of Natural Language Processing (NLP) resources and tools. In comparison, resources for Punjabi dialects, the informal spoken varieties of Punjabi, are still lacking. Punjabi dialects also present many challenges for conversion as one-to-many word mapping, incorrect Punjabi spellings and sub-dialects. In this paper, we present a conversion system for Punjabi text to its equivalent Punjabi dialect text which could be extended easily to be applied to other dialects of Punjabi. The proposed system is developed using the rule based approach that mainly relies on morphological analysis, conversion rules and bilingual dictionaries. Firstly, various Punjabi words, that need conversion, are identified. Then these identified words are replaced using a rule-based component that consists of conversion rules and bilingual dictionaries. A manual error analysis of system's outcome shows that it produces correct conversion over 95% for Malwai dialect and over 94% for Doabi dialect. The development of these rules is not an easy task as dialect form originates from pronunciation by native speakers.

Keywords: Bilingual Dictionary, Dialect Conversion System, Machine Translation, Punjabi Dialects

1. Introduction

Punjabi is an ancient and Indo-Aryan language which is mainly used in the Punjab region of India and Pakistan. It is one of the world's 14th widely spoken languages of the world. Punjabi is spoken by the inhabitants of countries namely India, Pakistan, Canada, England and America. Punjabi is having a variety of dialects, which are due to geographical locations and religious communities. The dialect is a variety of language that is different from other varieties of same language by features of phonology, grammar and vocabulary[1].A famous saying in Punjabi is that language in Punjab changes every half mile[2]. The main dialects of Punjabi are Majhi, Malwai, Doabi and Powadhi in India, and Lahndi, Pothohari and Multani in Pakistan. Majhi is the dialect used in both Amritsar and Lahore and is the standard written form of Punjabi. Indian languages are having lack of textual dialect processing and conversion tools as compared to foreign languages. As per the requirements, different dialect processing systems use various machine translation approaches. Most of the systems use the hybrid approach, which is a combination of rule-based approach and statistical approach, for better translation.

2. Need of Dialect Conversion System

There is lack of Punjabi dialect resources and NLP tools as compared with processing tools available for Standard Punjabi language. The Punjabi is spoken by more than 100 million people throughout the world, but still Punjabi has not risen to the status of a powerful language[3]. The

* *Author for correspondence*

demand of conversion systems become higher in past years due to increase in the communication between the various regional communities. Without these types of systems, the only feasible alternative is adoption of a single language which involves dominance of chosen language over other language. Since the loss of language involves disappearance of a distinctive culture. So there is a need to develop conversion systems which processes the regional languages. In most of the cases, speakers of one Punjabi dialect are sometimes unable to understand the some words of speech in other Punjabi dialect. The problem is due to different meanings of a particular Punjabi word in different dialects. As in case of "ਵਚਿਕਾਰ"(vIckar)' word the inhabitants of Malwa region use "ਦਰਮਿਆਨ"(dəmIan)' for the "ਵਚਿਕਾਰ"(vIckar)' word and residents of Doaba region use the 'ਗੱਭੇ(gəḅe)' word. This multiple meanings of Punjabi word acts as communication barrier when two individuals from different Punjabi spoken regions are interacting. The dialect acts as an identity maker for a particular community and geographical location[2]. So there is need to develop Punjabi dialect conversion systems to identify the particular dialects.

3. Linguistic Resources

There is no conversion system for Punjabi dialects. Our study is the first attempt to process the Punjabi dialects in a computational point of view, but many textual resources are available that are used for developing Punjabi dialect conversion system. Singh[4] has discussed the grammatical structure of Doabi and Malwai dialects. Author has given many suitable examples, which shows the clear difference between the Punjabi dialects and Standard Punjabi. A pilot study is done for collecting the Punjabi dialect data, mainly in Sangrur, Ferozepur, Ludhiana and Moga districts of Malwa region, and Nawanshahr and Kapurthala districts of Doaba region. Author has elaborated the syntactic and morphological structure differences between the phrases and sentences used in the Doabi and Malwai dialects[4]. A linguistic dictionary of Malwai dialect is developed by Singh[5]. The Malwai dialect is mainly influenced by the social and political factors, and geographic locations. Author has discussed various word level differences in the Malwai text from the Standard form of Punjabi language. The dictionary contains nearly 19000 words of Malwai dialect. The dictionary consists of Part - of - Speech tags related characteristics of the Malwai dialect words. Kumari[6] has discussed the Doabi dialect linguistic features, mainly of the Hoshiarpur district. Author has elaborated different syntactic and morphological structure characteristics of the words used in the Doabi dialect. There are very few cases in which one or more consonants are at starting position of the word. Different Part - of - Speech tags of Punjabi words, phrases and sentences are also discussed using suitable examples. Dialects play an important role for analyzing human psychology[7].

4. Methodology

The Punjabi dialect data is collected from various textual contents, available on personal blogs, social chats, discussion forums and Punjabi dialectology textual resources. After that, the data is analysed; Punjabi dialect data contents have been retained. By using these filtered out data contents bilingual dictionaries and morphological conversion rules are developed. The rule based component that consisting of Punjabi text to Punjabi dialect dictionaries and conversion rules is used for conversion task. The bilingual dictionaries are used for direct word-to-word conversion, whereas the conversion rules are used for replacing/removing specific portion of the input word. Figure 1 describes steps involved for the development of proposed system.

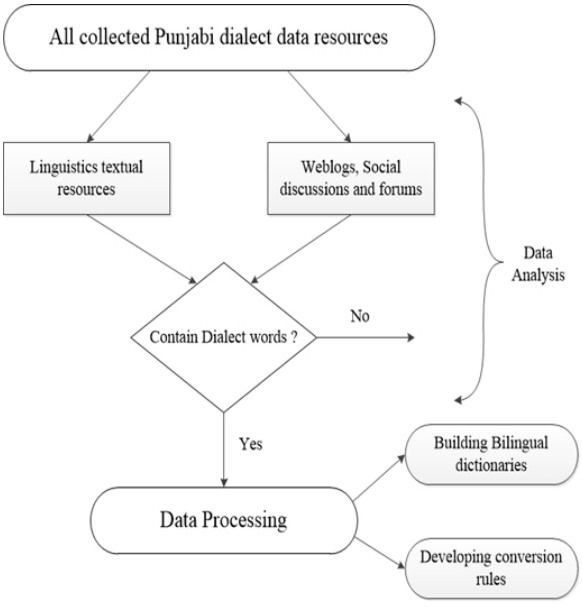

Figure 1. Flow Diagram of Development Process.

4.1 Data Resources Collection

There is lack of available resources in a particular Punjabi Dialect as compared to Standard Punjabi language resources and linguistic tools. The first task is to find instances of written Malwai and Doabi dialects, as Dialect form is more common in spoken form than in written form. The data is collected from various books, research papers, thesis, weblogs, online user groups and social discussions. The major resources of Punjabi dialects are Punjabi dialectology textual resources.

4.2 Data Analysis

This is actually the first step of the conversion process. Various Punjabi dialect resources, collected in the previous stage, are analysed. Those resources, which contain some number of Malwai and Doabi dialect words, are retained. The most likely Punjabi dialect words are manually filtered out from various resources namely weblogs, social discussions, forums and dialectology linguistic resources. The Punjabi dialect words are largely found on various Punjabi dialectology textual resources as compared to weblogs and social discussions. The filtered data is categorized in two categories as words in Malwai dialect or words in Doabi dialect.

4.3 Data Processing

This is main stage of development as various direct word-to-word mappings and conversion rules for converting Punjabi text to Punjabi Dialect conversion are developed. The filtered data has been processed and examined in the data processing phase. The filtered data has been used for developing the rule based component, which performs the conversion task. The rule based component consists of bilingual dictionaries and morphological conversion rules. The bilingual dictionaries are capable for performing direct word-to-word mapping. The conversion rules have developed for converting some specific portion of the input Punjabi word as first character, first vowel, last vowel and Part - of - Speech tags category. The proposed system is also capable for conversion of one Punjabi dialect to another using some conversion rules. The system training is performed using three dictionaries along with the general morphological conversion rule.

4.3.1 Building Bilingual Dictionaries

In order to develop the dictionaries, the filtered out Punjabi Dialect words are used. The system consists of three bilingual dictionaries, consisting of 1350 words of Standard Punjabi, Malwai dialect and Doabi dialect. Mostly the dictionary's entries are attained by the data available on Punjabi dialectology linguistic resources. Table 1 shows the bilingual dictionary sample, used for conversion to Malwai text.

Table 1. Standard Punjabi to Malwai Dialect bilingual dictionary sample

Standard Punjabi	Malwai Dialect
ਉਲਜਲੂਲ(uljəlul)	ਜਬਲ਼ੀ(jəbḷi)
ਬਰਾਬਰ(bərabər)	ਬਰੋਬਰ(bərobər)
ਭਾਈ(ḍ̣ai)	ਬਾਈ(bai)
ਨੂੰਹ(nũh)	ਨੋਹ(nôh)
ਚੁਗਲੀ(cUgli)	ਭਾਨੀ(ḍ̣ani)
.	.
.	.
.	.

The dictionaries are capable of handling the word level conversion and only used for direct word-to-word mapping. For example, 'ਬਰਾਬਰ(bərabər)'Standard Punjabi word is mapped to 'ਬਰੋਬਰ(bərobər)' *Malwai* dialect word. In the example direct word - to - word mapping is done as the whole word 'ਬਰਾਬਰ(bərabər)' is mapped to 'ਬਰੋਬਰ(bərobər)' word. In the same way the Standard Punjabi words are mapped to Doabi dialect words. The Malwai dialect to Doabi Dialect bilingual dictionary is used for inter dialect conversion.

4.3.2 Developing Morphological Conversion Rules

Some morphological conversion rules are developed which will convert the words left after direct word-to-word mapping. These rules are used for better translation. In some cases, specific portion of input words need to get converted. These rules are used for this purpose that generally converts the specific portion and Part - of - Speech (POS) tags category of the input Punjabi text. Some rules have been developed for converting the POS tags of the source text. These rules do not automatically identify the POS categories from the input text. Various

POS tag categories for Punjabi language are pronouns, verbs, adverbs, adjectives and prepositions. figure depicts the conversion of specific portion of the identified Punjabi word. Figure 2 the specific portion 'ਦੋ(dó)' of Punjabi word 'ਜਦੋ(jdó)' is converted to 'ਦੂੰ(dũ)' as the final converted word is 'ਜਦੂੰ(jdũ)'.

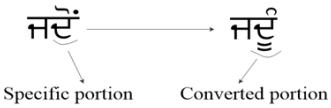

Figure 2. Conversion Rule for Replacing Last Two Characters Combination.

5. Proposed System Architecture

Figure 3 shows the flow diagram of main components of the conversion system. The proposed system has three main components namely Morphological Analyzer, Conversion Engine and Generator. Firstly Morphological Analyzer, by using different techniques, identifies various Punjabi words from the source sentence that need conversion. Conversion engine contains conversion rules and Malwai and Doabi dialect dictionaries that perform the conversion task. The conversion quality of system depends upon the efficiency of conversion rules and size of corpus. The Generator component produces the final outcome in desired dialect. The input and output of system is encoded in Unicode.

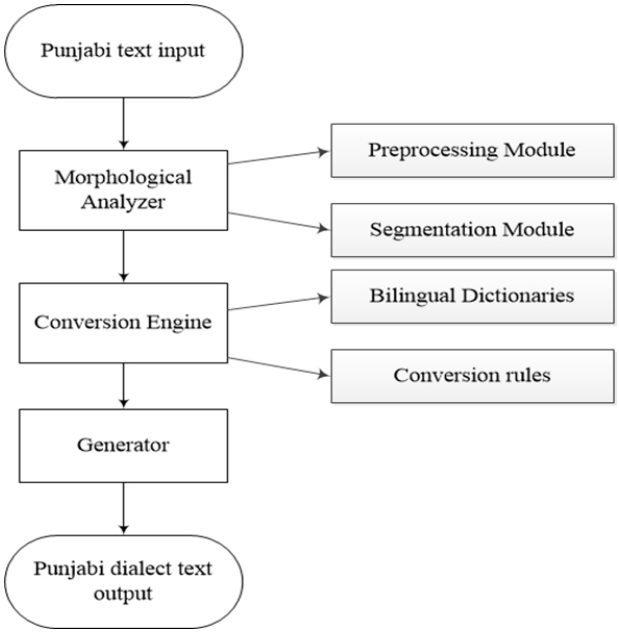

Figure 3. Flow Diagram of the Conversion System.

5.1 Morphological Analyser

This component identifies the words of the input text and processes the input text in such a way, that it matches with the data on which the system is trained on. This component examines the content of the source text to distinguish the words that need conversion. The component decides which words are to be converted and which are to be left.

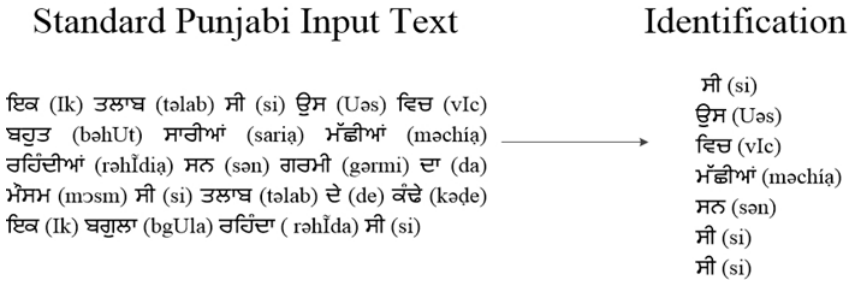

Figure 4. Identification of the Words From Input Text.

There are basically two modules of the analyzer namely prepressing module and segmentation module. These two modules are illustrated in following subsections.

5.1.1 Pre-Processing Module

This is the first module of the conversion system which gets Standard Punjabi text as input and analyses the word using some preprocessing steps. The main step of this module is to identify the various Standard Punjabi words that need the conversion. The words are identified by the various bilingual dictionaries and morphological conversion rules. After the identification of the words are passed to segmentation module for further processing. Various identified words for conversion to Malwai dialect are shown in the below figure 4.

5.1.2 Segmenting Module

The segmentation module divides the source text words into smaller units. The words are segmented based on the Unicode. The module helps the various conversion rules to convert the specific part of the input word as the input Punjabi word is a group of various vowels and consonants. In the following figure 5 the input Punjabi word "ਕਵਿੇਂ"(kIvė)' is segmented into 'ਕ(k)' 'ਿ(I)' 'ਵ(v)' 'ੇ(e)' and 'ਂ(·)' segments. The 'ਿ(I)' does not come at the first position of any Punjabi word. Using these segments the conversion rule converts the "ਕਵਿੇਂ"(kIvė)' word to "ਕਮਿੇ"(kImė)' word.

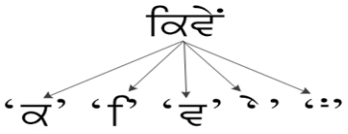

Figure 5. Segmentation of the Identified Punjabi Word.

5.2 Conversion Engine

Conversion Engine is the main component of proposed system. Various Punjabi source text words, which are identified by morphological analyzer, are converted to its equivalent Punjabi dialect words. The system training is done in two steps. In the first step, the selected words of input Punjabi text are mapped with its corresponding Punjabi dialect words using bilingual dictionaries. The words left after first step are processed by various conversion rules.

5.2.1 Word-to-Word Mapping

Three bilingual dictionaries are used for direct word-to-word mapping. The bilingual dictionaries contain words of Malwai dialect and Doabi dialect. For example 'ਭਾਈ(ḅai)' word is mapped to 'ਬਾਈ(bai)' word. In the example direct word-to-word mapping is done as the whole word 'ਭਾਈ(ḅai)' is mapped to 'ਬਾਈ(bai)'equivalent Malwai dialect word.

5.2.2 Using Morphological Conversion Rules

Conversion rules are used for converting the words left after processing by bilingual dictionaries. These rules are used to replace specific portion of source text. Some rules have been developed that convert Part - of - Speech tag category of the input text. There are some rules which map the first vowel, first consonant and last vowel to the target text. Figure 6 shows the conversion by first character rule.

In the figure 'ਵ(v)', the first consonant of input word is converted to 'ਬ(b)'.

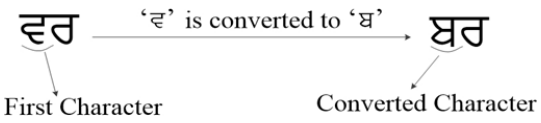

Figure 6. Conversion Rule for Replacing First Consonant.

5.3 Generator

After the conversion engine has converted all the identified words, the generator component generates the output. The component uses the previous analyses to generate different Punjabi dialect language words. The component replaces each unit identified above with the corresponding converted word of the desired dialect. After that, the generator produces the final outcome in the desired dialect. The output text is Unicode encoded. The outcome accuracy highly depends upon the size of training data.

6. Evaluation

Text collected from different literature resources have been used for evaluation purpose. First of all direct word-to-word mapping of identified words is done using bilingual dictionaries and remaining words are converted by morphological conversion rules that match the character combination of the identified words. Word Accuracy Rate (WAR) is used for the evaluation of the conversion system. WAR is the percentage of correct conversion from the total generated conversion by the system. The metrics is used to measure the performance of the system at word level. Our system is tested over news data, short stories and Punjabi novels, consisting of 12000 words of Punjabi language. For Standard Punjabi to Malwai Punjabi conversion word accuracy rate is found to be 95% and word accuracy rate for Standard Punjabi to Doabi Punjabi has been found to be 94% as shown in table 2.

Table 2. Accuracy of the conversion system

Punjabi Dialect	Conversion Units (words)			Accuracy Rate (%)
	Total words (Input)	Wrong conversion	Right conversion	
Malwai	12000	610	11390	95
Doabi	12000	650	11350	94

The system also performs the task of inter dialect conversion as Malwai to Doabi conversion and Doabi to Malwai conversion. The inter dialect conversion output has been evaluated by comparing it with the output of the Standard Punjabi to dialectal Punjabi conversion. The accuracy of the conversion for Doabi dialect is less as there is lack of resources of the Doabi dialect. It seems to be extremely hard to develop a system which can truly give 100% accuracy. Wrong conversions mainly occurred due to spelling errors, proper nouns and mismatching of conversion rules. The results are quite promising and show the success of first dialect conversion system for Punjabi language.

7. Conclusion

The proposed system has been designed and implemented for conversion of text written in Punjabi text into its Malwai dialect and Doabi dialect equivalents using rule based approach. The proposed system identifies the words from the input text, splits these into individual segments and then converts them to its equivalent Punjabi Dialect text. The main component of proposed system is conversion engine that relies on bilingual dictionaries and morphological conversion rules. Only a single rule is not applicable for all the words so different rules have been proposed for conversion as for POS tag categories. Various rules are also proposed to replace first character, last vowel and last vowel and consonant combination. The development of these rules is not an easy task as dialect form originates from pronunciation by native speakers and there are very less linguistic resources available from which the conversion rules are developed. Better results are achieved by increasing the size of the training data. The results are quite promising and show the success of first dialect conversion system for Punjabi language. This is the first conversion system for Malwai and Doabi dialects.

The future work will be based on using various other Machine translation approaches. Also the coverage of system in the handled dialects and to new dialects can be extended. The system can further be improved to automatically learn new morphological conversion rules from limited available data. In future there should be development of conversion systems for converting Punjabi dialects to other languages as Hindi and English.

8. References

1. Marimuthu K, Devi SL. Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil Text using FSTs. In: Joint Meeting of SIGMORPHON and SIGFSM; 2014 Jun; Maryland, USA. p.37-45.
2. John A. Two Dialects One Region: A Sociolinguistic Approach to Dialects as Identity Markers [Master Thesis]. Indiana: Ball State University; 2009.
3. Gillani M, Mahmood MA. Punjabi: A Tolerated Language Young generations' attitude. Research on Humanities and Social Sciences. 2014; 4(5):129-137.
4. Singh H. A comparative study of Doabi and Malwai dialects [MPhil Thesis]. Patiala: Punjabi University; 2007.
5. Singh M. Malwai Kosh. Patiala: Publication Bureau Punjabi University; 2007.
6. Kumari R. Dawabi da up-bhashai sarvekhan Hoshiapur zile de parsang wich [PhD Thesis]. Chandigarh: Punjab University; 2002.
7. Singh P. Sindantak Bhasha Vigyan. Patiala: Madan Publishers; 2010.