International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Standard Punjabi Text to Lahndi Dialect Text Conversion System

Parneet Kaur[1], Simrat Kaur[2]

[1, 2]Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib
[1]Research Scholar, [2]Assistant Professor

Abstract: ***In modern world, web is used as a medium to exchange information, ideas, and thoughts etc. Further people like to exchange this information in their native language. There are many online translators which translates one language to another or from many languages to English, so that everyone can get the required information even if one does not understand that language. In recent times people has started using colloquial data also. Translation of colloquial and informal data presents many challenges as one-to-many word mapping, sub dialects and incorrect spellings. In this paper, a system is presented for the conversion of Punjabi text to its Dialectal Lahndi form. The system is developed using a rule based approach which include approximately eighty five rules for conversion and a bilingual dictionary consisting 3247 words. The conversion system first segments the sentences into words and identifies the words which need conversion and then it converts the words by applying conversion rules and bilingual dictionary. This Conversion system can be extended to other dialects of Punjabi also.***

Keywords: Rule based approach, Machine Translation, Bilingual Dictionary, Punjabi Dialects, Colloquial data.

1. Introduction

Punjabi being one of the world"s 14 widely spoken languages, is a language of more than 100 million people throughout the world [3]. It is mainly spoken in eastern and western Punjab region. Outside India and Pakistan it is also the language of inhabitants of Canada, England, America. The language used in most of the Punjabi books, official work, literature, education, art etc, is the Standard Punjabi. Majhi, a dialect of Punjabi is considered as the base for Standard Punjabi [5]. Punjabi is the only language that is completely tonal [1]. Due to geographical locations and religious communities, a range of informal and dialectal forms of the Punjabi language have emerged [2]. Many scholors has classified Punjabi into its dialects and their classification differs from one another. Grierson divided Punjabi into Eastern Punjabi and Western Punjabi or Lahndi. The main dialects of Punjabi are Majhi, Malwai, Doabi and Powadhi in India and Lahndi, Pothohari and multani in Pakistan [1]. Lahndi is a group of languages in western Punjab [8]. It is spoken in area of Atak, Rawalpindi, Jehlam, Shahpur, Miawali, Mujafargarh, Multan, jhang, Saiwal, Faislabad, Bannu, Dera Ismaeel khan, dera Gaji Khan. There are many language processing tools for regional languages but lacking dialectal processing tools. Punjabi Dialect processing tool development is a challenging task as it presents many challenges as dialects contain many words for a same word in standard Punjabi or different spelling for same words.

2. The Need of Conversion System for Dialects

Punjabi is official language of Indian state of Punjab and also one of the official languages of Delhi [6]. In India there are more than 50 recognized languages in the world and Punjabi is very important among these Indian languages [12]. Punjabi has got a status in India but being a language of more than 100 million people, it has not been risen to a powerful status. Also the speakers of Punjabi are more in Pakistan state of Punjab i.e., Western Punjab than Eastern Punjab. Even though Punjabi has become a tolerated language in Western Punjab. People are shifting their language form Punjabi to Urdu [10]. Further Lahndi is also mainly spoken in Western Punjab and most of the literature from times of Baba Farid to Maharaja Ranjit Sing is written in Lahndi Dialect. To raise the status of Punjabi and to identify between its dialects Dialectal processing tools are required. Also communication through web is the need of modern world. If these type of systems does not exist then translation between regional languages can be done but translation between dialects of a language cannot be done, hence it will require adoption of a single dialect out of many dialects which dominate the other dialects and cause their extinction. So Dialectal Conversion systems are the need of today"s world. Also these remove the barriers of communication as sometimes speakers of one dialect cannot understand some words of another dialect because of its different meaning in another dialect. For example for the Standard word 'ਵਿਚਕਾਰ' speakers of Lahndi use „ਗੱਭੇ" and speakers of Malwai use „ਦਰਮਿਆਨ".

3. Literature Survey

There are Dialectal Conversion Systems for Arabic dialects but in Indian languages, very less work has been done.

Marimuthu, K. and Devi, S. L. (2014)
The system can translate various spoken Tamil dialects to Standard Written Tamil text. Finite State Transducers are used for obtaining equivalent Standard Tamil words and Conditional Random Fields are used for handling agglutination and compounding in the resultant text. The system can translate central Tamil, Madurai Tamil, Tirunelveli Tamil, Brahmin tamil, kongu Tamil and common spoken forms. The translation accuracy is higher for Kongu Tamil dialect and lower for Madurai and Tirunelveli due to polysemous nature of the words of these dialects [11].

Singh, A. and Singh, P. (2015)

The system translates sentences between two dialects of Punjabi language-Malwai and Doabi, and from Standard Punjabi to these dialects. A Rule base approach is used with three bilingual dictionaries that translates Standard Punjabi to Malwai, Standard Punjabi to Doabi, Malwai to Doabi and Doabi to Malwai. Accuracy of the system for Standard Punjabi to Malwai is 95% and Standard Punjabi to Doabi is 94% [1].

Sawaf, H. (2010)
The system is an extension of a Hybrid Machine Translation System for handling Arabic dialects. It uses a Statistical decoder which contains four types of rules-lexical, syntactic, argument structure, and functional structure rules, semantic disambiguation information, a statistical bilingual lexicon, bilingual phrase table and target language models. The system is tested with and without dialect normalization against BLEU score and result is higher score with dialect normalization [7].

Sarath, K. S., et al. (2014)
The system translates informal sentences and slangs of Thrissur dialect to a formal format. A hybrid approach mixing Rule bas and machine learning approaches is used. Accuracy in the following target words are depends upon the previous resolved formal words [9].

Salloum, W., et al. (2012)
This system improves the output of different previously developed MT systems by selecting what sentence go to which MT system. The system consider two dialects-Levantine and Egyptian along with MSA. This system can identify the type of sentence if it is a MSA only sentence or include some dialectal content so that corresponding suitable MT system can be used and the accuracy for the output increases. This best system selection approach improves over the best baseline single MT system by 1.0% absolute BLEU point on a blind test set [13].

Zaidan, F. O., Burch, C. C. (2011)
This system holds 52M –word monolingual dataset which is rich in dialectal content. Also the system is trained to identify the dialectal content and to specify the level of dialectal content in a sentence. The data is extracted from three newspapers which contained high degree of dialectal content from Levantine, gulf and Egyptian dialects. The system can distinguish the dialectal content from MSA and from other dialectal content [4].

4. Methodology

For the collection of dialectal data, various Punjabi dialectology textual resources, personal blogs and social chats are studied. Then data analysis is done to identify the words which are frequently used and different from Standard Punjabi words. The analysed data is then processed to make bilingual dictionary and conversion rules. Bilingual dictionary is used for word-to-word mapping and conversion rules are used for translation of those words whose certain portion matches the Punjabi dialectal word, for example 'ਬੁਣਨਾ' is a Standard Punjabi word which is replaced with Lahndi dialect word 'ਉਣਨਾ' using conversion rules.

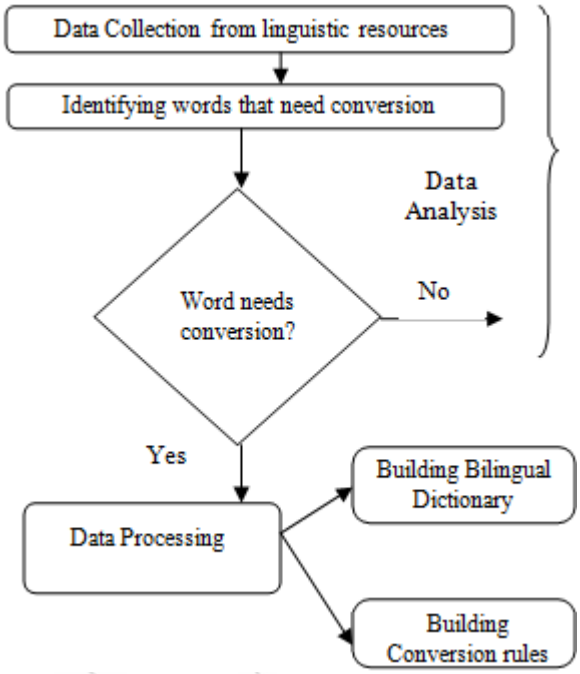

Figure 1: Flow Diagram of Development Process

4.1 Data Collection

Due to lack of linguistic resources for dialectal content, the first task is to find the resources, books, social chats, blogs, dictionaries which hold the Lahndi dialectal content. Most of the words are taken from a Lahndi dictionary. Other resources are studied to find most frequently used words.

4.2 Data Analysis

The data which is collected in previous step is analysed to identify most frequently used words that need conversion. One-to-many word mappings are resolved manually in this step by choosing a single appropriate word out of many words. Data Analysis is considered as first manual step for conversion process.

4.3 Data Processing

Data Processing is the main stage of development process. In this stage bilingual dictionary and conversion rules are developed for translation of Standard Punjabi Text to Lahndi dialect text. The filtered data in the previous stage is used to develop a rule based component which consists a bilingual diactionary and conversion rules for translation. Bilingual dictionary is developed for direct word-to-word mapping and conversion are developed for replacing/removing certain portions of a word.

4.3.1 Building Bilingual Dictionary

Analysed data after removing one-to-many word mappings is used to build bilingual dictionary. The dictionary is used for direct word-to-word mapping. It consists of 3247 words of Standard Punjabi. Corresponding to Standard Punjabi word, there is Lahndi dialect word. When a word from input string

matches a Standard Punjabi word in bilingual dictionary, whole word is replaced with the corresponding Lahndi word. Table 1 shows a sample of Bilingual dictionary.

Table 1: Bilingual Dictionary Sample

Standard Word	*Lahndi Word*
ਨੇੜੇ	ਲਵੇ
ਹਾਏ	ਉਈ
ਉਠ	ਸ਼ੁਤਰ
ਉਦਾਸ	ਉਚਾਟ
ਪੰਛੀ	ਪੰਖੀ
ਇਤਰਾਜ਼	ਉਜ਼ਰ

Bilingual dictionary is capable of only direct word-to-word mapping. For example Standard Punjabi word „ਨੇੜੇ" is mapped to Lahndi dialect word „ਲਵੇ". The direct word-to-word mapping is done as a whole word.

4.3.2 Developing Conversion Rules

There are 85 Conversion rules for translating certain portions of words. For example in Figure 2 the Standard Punjabi word „ਔਲਾਦ" is converted to „ਉਲਾਦ" by replacing „ਔ" by „ਉ".

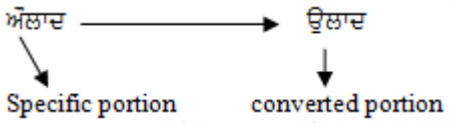

Figure 2: Conversion Rule for replacing first character

The conversion rules are general as they are applied not only on one word but on all those word which satisfy the rule. There are also rules for POS tag categories. The system does not detect the POS automatically but translate POS tag category words if they satisfy a rule.

5. System Architecture

Figure 3 shows a flow diagram of main components of the Standard Punjabi to Lahndi dialect conversion system. The system has three main components- Morphological Analyzer, Conversion Engine and Generator. First two components- Morphological Analyzer and Conversion Engine work in a combined way. First Standard Punjabi text is input to the system. Morphological Analyzer divide the input strings into words separated by commas, then Conversion Engine by considering the Bilingual Dictionary, replace the words that match in dictionary. Then again for remaining words Morphological Analyzer do segmentation of words. Rules are applied to these segmented words and replaced if they satisfy the rule. The Generator at the end generates output. The input and output of the conversion system is encoded in Unicode.

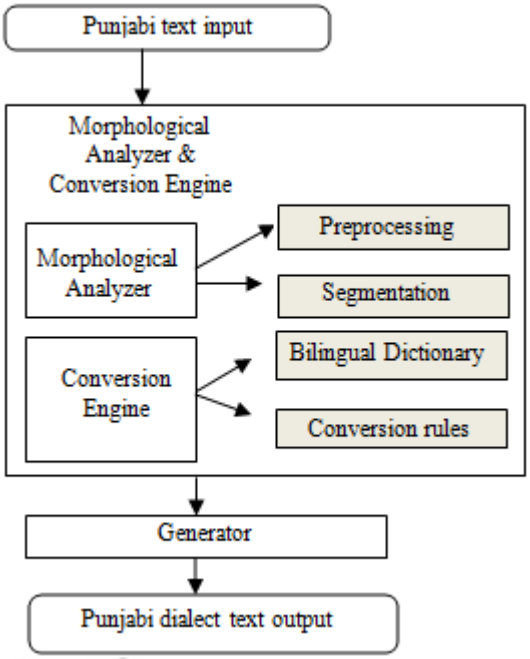

Figure 3: Architecture of Conversion System

5.1 Morphological Analyzer and Conversion Engine

In this component Morphological Analyzer and Conversion Engine work interchangeably. First Morphological Analyzer split the input string of Standard Punjabi words into independent words and then identify the words those matches the words in Bilingual dictionary. After identification Conversion Engine replace those words with corresponding Lahndi words. The words once replaced here are not applicable for rules to be applied. Morphological Analyzer segments the remaining word into characters. Again Conversion Engine checks each word for a rule to match. Once a rule applicable on the word is found, the word is replaced.

5.1.1 Morphological Analyzer

It consists of two modules. The main task of this component is to identify those words which need conversion and to segment the words left after word-to-word mapping.

5.1.1.1 Preprocessing module

This module divides the input text into independent words and the compare each word with the words in the Bilingual Dictionary. If the word matches, the module identifies it as the word need conversion.

5.1.1.2 Segmentation module

After word-to-word conversion using Bilingual Dictonary, this module segment the remaining words into characters or smaller units. The word segmentation is based on Unicode. This module helps Conversion Engine to apply rules as rules are based on these segmented smaller units. For example in Figure 4 the word „ਇਕੱਠ" is segmented to „ਇ", „ਕ", „ੱ", „ਠ" and replace the first segment with „ਅ" to generate word „ਅਕੱਠ".

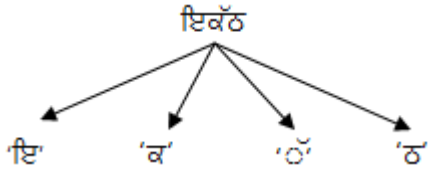

Figure 4: Segmentation of the remaining word

5.1.2 Conversion Engine

Conversion Engine is the main component of the Conversion System. It converts the words in two steps. In first step the words identified by Preprocessing module are mapped with corresponding Lahndi dialect words using Bilingual Dictionary. In second step the words processed by Segmentation module are converted using conversion rules.

5.1.2.1 Word-to-Word Mapping

In word-to-word mapping whole word is replaced by dialectal word. Bilingual Dictionary is used for this task. For example the Punjabi word ,ਨੇੜੇ" is mapped to Lahndi dialect word ,ਲਵੇ".

5.1.2.2 Using Conversion Rules

Conversion rules are used for replacing or removing certain portion of a word. Only the words processed by Segmentation module are eligible for rules. If two rules are applicable on a single word, then only one rule which comes first modify the word and second rule cannot modify the word. For example Figure 5 shows a conversion using first character rule. The first consonant ,ਦ" is replaced by ,ਡ".

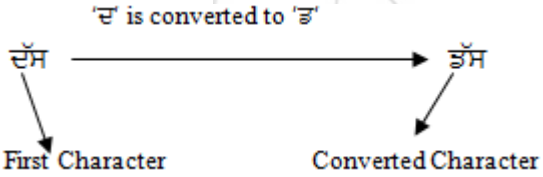

Figure 5: Application of Conversion Rule

5.2 Generator

After conversion of the matched words, Generator generates the output in a Unicode encoded text. The generator organize the replaced and converted words and show the output again in strings of words.

6. Evaluation

Word Accuracy Rate (WAR) is used for the evaluation of the conversion system. Words are first converted using bilingual dictionary and morphological conversion rules are applied on remaining words. WAR is the metric used to measure the performance of the system at word level. WAR is the percentage of correctly converted words to the total generated conversion by the system. The system is tested over news data, novels, stories, consisting of words of Punjabi language. The Word Accuracy Rate for the system is .

Table 2: Accuracy of the Conversion System

Total Words (Input)	Right Conversion	Wrong Conversion	Accuracy Rate (%)
8460	8179	281	96.7

Wrong conversions by the system are generated due to spelling errors, mismatch of rules and proper nouns.

7. Conclusion and Future Scope

The system proposed give promising results with Word Accuracy Rate of . The system is based on Rule based approach. The conversion process is done by two main components Morphological Analyzer that identifies and segments the words and Conversion Engine that converts the words first by using Bilingual Dictionary and then by applying Morphological rules on remaining words. Accuracy of the system relies on the size of training data and conversion rules. The system consists of 3247 words in Bilingual Dictionary and 85 rules.

In future more rules can be added to increase the accuracy of the system and also the system can be developed with other Machine Translation approaches. The system can be extended to other dialects of Punjabi. The conversion system can be trained to learn more morphological rules by itself in future. The system will be helpful in translation of Punjabi dialects to Hindi or other International languages.

References

[1] Singh, A. and Singh, P., "Punjabi Dialects Conversion System For malwai and Doabi Dialects", Indian Journal of Science and Technology, ISSN: 0974-6846, Vol. 8, Issue 27, pp 1-6, October 2015.
[2] Singh, A. and Singh, P., "A Rule Based Punjabi Dialect Conversion System", International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653, Volume 3 Issue VI, June 2015.
[3] Asher, J., "Two Dialects One Region: A Sociolinguistic Approach to Dialects as Identity Markers", Thesis, Ball State University, Muncie, Indiana, 2009.
[4] Zaidan, F. O., Burch, C. C., "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, Portland, Oregon, pp 37–41, June 19-24, 2011.
[5] Kaur, G., "Punjabi bhasha da taksalikaran", Thesis, Guru Nanak Dev University, Amritsar, 2007.
[6] Josan, G. S. and Lehal, G. S., "A Punjabi To Hindi Machine Translation System" Coling 2008: Companion volume-Posters and Demonstrations, pp 157-160, August 2008.
[7] Sawaf, H., "Arabic Dialect Handling in Hybrid machine Translation", In Proceedings of the Conference of the Association of machine Translation in the Americas(AMTA), Denver, Colorado.
[8] Bahri, H., "Lahndi Kosh", Anand Sons, New Delhi, Dr. Param Bakshis Singh, Punjabi Univ.ersity Patiala, 2005.

[9] Sarath, K. S., et al., "Dialect resolution: A Hybrid Approach", An International Journal of Engineering Sciences, Special Issue iDravadian, Vol. 15 ISSN: 2229-6913, December 2014.
[10] Gillani, M. and Mahmood, M.A., "Punjabi: A Tolerated Language Young generations‟ attitude", Research on Humanities and Social Sciences, ISSN (Paper) 2224-5766, Volume 4, Issue 5, pp. 129-137, 2014.
[11] Marimuthu, K. and Devi, S. L., "Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil Text using FSTs", Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA, pp 37-45, June 27 2014.
[12] Singla, S. and Baghla, S., "Hybrid Approach for English to Punjabi Translation System for Newspaper Headlines in a Specific Domain", International Journal of Engineering Research and Technology, ISSN 2278-0181, Vol. 2, Issue 11, pp 1792-17995, November 2013.
[13] Salloum, W., et al., "Sentence level Dialect Identification for machine translation System Selection", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, Maryland, USA, pages 772–778, June 23-25 2014.

Author Profile

Parneet kaur received the B tech. degree in Computer Science and Engineering from Guru Gobind Singh College of Modern Technology in 2014. She now pusuing her M. Tech. from Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib.

Simrat Kaur received B tech. degree in Computer Science and Engineering from RGPV, Bhopal in 2002 and done her masters degree in 2007 from Guru Nanak Dev Engineering College, Ludhiana. She is now pursuing her Phd. Her area of specialization is Natural Language Processing. She has 13 years of teaching experience and presently serving BBSBEC, Fatehgarh Sahib. Her 6 papers are published in International Journals and 3 in National Conferences.