iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 3 **Issue:** VI **Month of publication:** June 2015

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue VI, June 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# A Rule Based Punjabi Dialect Conversion System

Arvinder Singh[1], Parminder Singh[2]
[1]M.Tech. Student, [2]Associate Professor
CSE Department, Guru Nanak Dev Engineering College, Ludhiana

***Abstract — Recently there is dramatically increase in availability of informal dialectal language on web. People are sharing their ideas, thoughts and emotions, usually in the form of blogs and partially informal articles across the www. We present a system that will translate standard Punjabi text to given Punjabi dialect. There is a lack of language processing tools of dialectal Punjabi language in comparison with standard Punjabi language processing tools. Our study describes the problem and need for processing Punjabi dialect language. The developed system employs a rule based approach which processes only Malwai dialect and Doabi dialect of Punjabi. Different techniques are used to identify standard words in a source. Selected words are replaced using a rule-based component, which contains transfer rules and Standard Punjabi-Dialectal Punjabi dictionaries.***
***Keywords — Punjabi, Punjabi Dialects, Bilingual dictionaries, Morphological transfer rules, Conversion engine.***

## I. INTRODUCTION

Punjabi is an Indo-Aryan language spoken by 102 million speakers worldwide, including countries India, Pakistan, Canada, England and America. Due to the geographical locations and religious communities, a range of informal and dialectal forms of the language have resulted. The dialect is described as a variety of language that is different from other varieties of same language by features of phonology, grammar and vocabulary [9]. A famous saying in Punjabi is that language in Punjab changes every half mile [1]. The dialects of Punjabi language differ widely among themselves, due to the socio-economic conditions of the speakers and depending on the geographic distribution. The main dialects of Punjabi are *Majhi*, *Malwai*, *Doabi*, *Powadhi*, *Multani*, and *Podhohari*.

There is a lack of dialect conversion system for Indian languages as compared to foreign languages. Various developed dialect conversion systems use different machine translation approaches as per the requirement. Most of the dialect processing systems use the hybrid approach, consisting of rule-based approach and statistical approach for better translation.

The rest of this paper is structured as follows: Section 2 explains the need of dialect processing and translation system. Section 3 discusses some of the challenges associated with processing Punjabi and its dialects. Section 4 discusses various linguistic resources of Punjabi dialectology and previous dialect conversion systems. Section 5 describes the steps for development of translation system. Section 6 discusses the main components of the developed dialect translation process. Section 7 shows results of developed system.

## II. NEED FOR DIALECT CONVERSION SYSTEM

There is lack of Dialectal Punjabi language processing 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 [11]. The demand of translation system becomes higher in past years due to increase in the communication between the various regional communities. Without machine translation 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 machine translation system is necessary for communication purposes between different regional communities.

## III. CHALLENGES IN DIALECT PUNJABI PROCESSING

The challenges present in for processing standard Punjabi language are also faced by dialectal Punjabi language. However, new challenges are also posed due to the lack of standard orthographies and informal spoken form of dialects.

*A. Lack of Linguistic Resources*

Dialectal Punjabi is impoverished in terms of resources and available tools, compared to Standard Punjabi. There are almost no Standard Punjabi (SP) to Dialectal Punjabi (DP) parallel corpora and very little parallel DP to English corpora. Most of the dialectal text is in spoken form that is informal and colloquial rather than standard written form.

*B. Unclear language boundaries*

Dialects are identity makers for a particular community, religion and geographic location but there is no clear DP language boundary for two adjacent dialects. This creates problem while identifying a particular dialect.

*C. One-to-Many word mapping*

The speakers of one Punjabi dialect are sometimes unable to understand the speech of other Punjabi dialect. The problem is due to one to many word mapping that is present in most of the cases. It is also very likely that multiple words within a dialect may map to a single word or more words in other dialect. Fig. 1 depicts one to many word mapping present in Punjabi Dialects as word ‘A`gy’ is ‘gwVI’ in *Majhi* dialect and ‘mUhry’ in *Malwai* dialect.

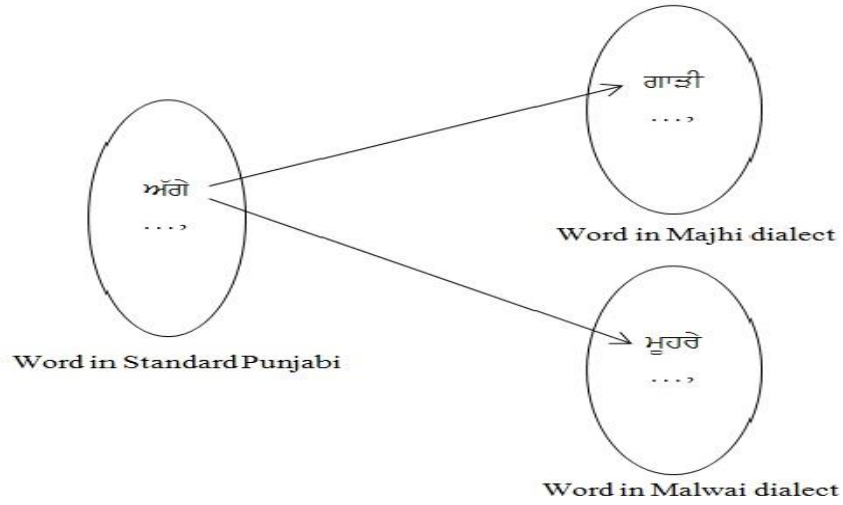



Fig. 1 An example of One-to-Many word mapping

*D. Mergence with other major languages*

Modern Dialectal Punjabi language gets merges with the Hindi dialects in northern India and with the Sindhi dialects in southern Pakistan. This colloquial language is also influenced by languages previously spoken in the areas. The presence of Urdu adds yet another layer of complexity.

*E. Spoken form of dialects*

Dialectal Punjabi words have inconsistent spellings. This problem is due to spoken form of dialects. Dialectal words are mainly used in spoken communication and when they are written by users, they do not conform to SP spellings.

## IV. RELATED WORK

*A. Linguistic Resources*

There is no conversion system for Dialectal Punjabi language. Our study is the first attempt to convert DP language in a computational point of view, but many textual resources are available that are used for developing Dialectal Punjabi processing system. Kaur focuses mainly on the phonetic decomposition of the dialectal words [10]. Author concentrates on the various

word level changes in the *Malwai* dialect words from the standard Punjabi. Some grammatical rules for the conversion of Punjabi text from *Malwai* dialect to the standard Punjabi are also described.

Kumari elaborates the various syntactic and morphological structure of the words used in the *Doabi* language. Author introduces various grammatical rules for the conversion of words from *Doabi* dialect to the standard Punjabi. The main focus is on the language of Hoshiarpur district [13]. Singh describes the grammatical and morphological structure of various words covered in *Malwai* dialect and *Poadi* dialect. Author mainly concentrates on the phonetic study of the words of the *Malwai* dialect [7]. Singh compares the grammatical structure of various words covered in *Doabi* dialect and *Malwai* dialect [5]. The main focus is on the part of speech tags of both the dialects. Author has elaborated various syntactic and morphological structure differences between the words used in the *Doabi* and *Malwai* dialect. Various speech related changes of *Doabi* dialect and *Malwai* dialect are also discussed.

### *B. Dialect Translation Systems*

Many conversion systems are available for converting and processing dialects, but there are very few system developed for Indian dialects. Marimuthu and Devi have presented a system which automatically converts Tamil dialect language text to standard written language [9]. The system is the first attempt to transform various dialects of Tamil to standard Tamil text. They are discussed various challenges and problems which are encountered while designing and developing the system. The conversion system has three main components as FST, CRF word boundary identifier and Sandhi Corrector. The FST is used for replacing the dialect language text suffixes with written language suffixes. The CRF word boundary identifier module specifies word boundaries in compounded words and breaks them into a set of simpler words. This identifier makes the morphological process much easier. Sandhi corrector is used for spelling checking. It makes necessary spelling changes and makes them meaningful and sensible simpler words. The system outcome undergoes two evaluation as direct evaluation and indirect evaluation. The direct evaluation is done using gold standards in which evaluation of system is done under precision, recall and F-measure. The performance of the system is higher under *Kongu Tamil* dialect as compared to other dialects because words of this dialect are rarely polysemous in nature.

Habash and Rambow have presented MAGEAD, a morphological analyzer and generator for the *Arabic* language family [12]. Author addresses the need for processing the morphology of the dialects. The proposed system provides an analysis of (root+pattern+features) representation. It has separate phonological and orthographic representations and allows for combining morphemes from different dialects. The evaluation is done using two strategies: a test suite of selected surface word/analysis pairs that test the breadth of phenomena covered, and a test corpus, which tests the adequacy on real text. The test suite is used for regression testing during development, as well as for qualitative assessment of the analyzer or generator. The only test corpus is Penn Arabic Treebank for MSA.

Bakr et al. have developed a machine translation system for converting written Egyptian colloquial sentences into Modern Standard Arabic language. The system uses hybrid approach along with Support Vector Machine approach for the diacritization of Arabic text. The first step of system is building colloquial lexicon and colloquial training corpus which consists of 41705 words. The next step is building training and test files. Then, using yamCha-0.33 tool, training and testing is performed. After training conversion of Colloquial word to corresponding MSA word is done. The results are promising and could be used with other colloquial languages. For further improving accuracy POS taggers are used [4].

Sawaf has proposed hybrid machine translation system for translating Dialectal Arabic to Standard Arabic. The system take advantages of both Rule based approach and Statistical approach. The system uses a decoding algorithm that normalizes non-standard and dialectal Arabic into Modern Standard Arabic. The translation quality is further improved by 2% for web text and 1% for broadcast transmissions by training [6].

Salloum and Habash have presented a rule-based approach for producing MSA paraphrases of dialectal Arabic words for OOVs and low frequency words. The approach extends an existing MSA analyzer and uses transfer rules to generate paraphrase lattices that are input to a phrase based statistical machine translation system. The system improves the quality of Arabic-English machine translation on dialectal Arabic text using morphological knowledge. The system accuracy claimed to be 74% for OOVs and 60% for low frequency words [17].

Zbib et al. have proposed a system for translating Dialectal Arabic to English language [15]. The System contains parallel corpus, consisting of 1.5M words. The dialectal sentences are collected from large corpus of Arabic web text and are translated using Amazon' Mechanical Turk. Sentence boundaries are important for correct translation; author segmented passages into

individual sentences using MTurk. These dialectal sentences are used to build Dialectal Arabic MT system. The system uses GIZA++ to align sentences and extract hierarchical rules. The decoder component use a log-linear model that combines the scores of multiple feature scores, including translation probabilities, smoothed lexical probabilities, a dependency tree language model, in addition to a trigram English language model. A set of experiments are performed on proposed systems trained using our dialectal parallel corpus with systems trained on a MSA-English parallel corpus. All experiments use the same methods for training, decoding and parameter tuning, and we only varied the corpora used for training, tuning and testing. The system claimed accuracy of 6.3 and 7.0 BLEU points higher than Modern Standard Arabic machine translation system on 150M word Arabic-English parallel corpus.

## V. METHODOLOGY

The data resources are collected from various textual contents, available on personal blogs, social chats, discussion forums and Punjabi dialectology textual resources. After that, the data is analysed; useful data contents are retained. The filtered data contents are used to build the bilingual dictionaries and morphological transfer rules. Using rule based component, consisting of transfer rules and SP-DP corpus, dialectal Punjabi words are formed. Fig. 2 describes steps involved for development of the proposed system.

### A. *Data Resources Collection*

There is lack of available resources of Dialectal Punjabi language in comparison with Standard Punjabi language resources. The first task is to find instances of written Dialectal Punjabi, as DP language is more common in spoken form than in written form. We have collected DP text from various books, research papers, thesis, weblogs, online user groups and social discussions. The major resources of Punjabi dialects contents are Punjabi dialectology textual resources.

### B. *Data Analysis*

This is actually the first phase of the machine translation process. Identifying the dialect of a text can be challenging in the absence of phonetic cues [15]. We analyze the various Dialect Punjabi resources, collected in the previous stage. Resources containing some number of dialect words are retained. The dialectal words are largely found on various Punjabi dialectology textual resources as compared to weblogs and social discussions. The most likely dialectal words are manually filtered out from various resources. We have classified the filtered data for being as in *Malwai* dialect and *Doabi* dialect.

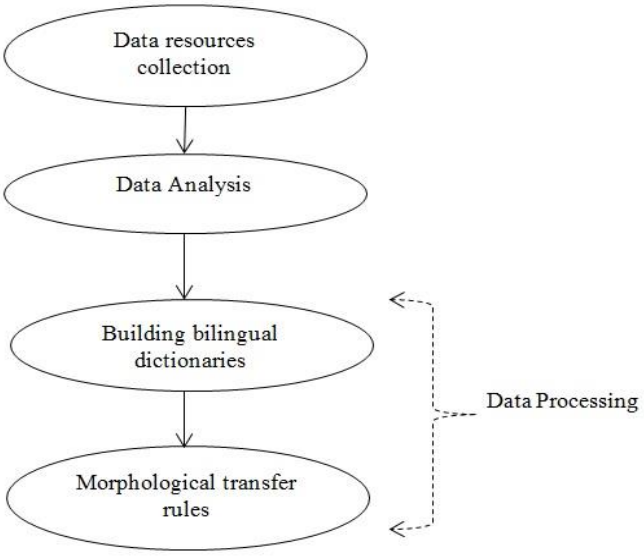



Fig. 2 Flow outline of various steps of development

C. *Data Processing*

The filtered data are processed and examined. Using the filtered data a rule based component is developed that contains bilingual dictionaries and morphological transfer rules. This component performs Standard Punjabi to Dialectal Punjabi conversion. The system training is performed using two dictionaries along with the general morphological transfer rules.

1) *Building Bilingual dictionaries*

In order to prepare the dictionaries, we use the filtered out dialectal words. For this system two dictionaries have been developed, consisting of 1150 words of *Malwai* dialect and *Doabi* dialect. Mostly the dictionaries entries are acquired with data available on Punjabi linguistic resources. The dictionaries are capable of handling the word level conversion and only used for word to word mapping. For example 'ਭੱਜ' word is mapped to 'ਨੱਠ' word. In the example direct word to word mapping is done as the whole word 'ਭੱਜ' word is mapped to 'ਨੱਠ' word.

2) *Developing Morphological Transfer rules*

Some morphological transfer rules are developed that will work along with dictionaries 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 POS tags portion of the input Standard Punjabi text. For example, 'ਕਿਵੇਂ' word is converted to 'ਕਿਮੇਂ' word. In the above example 'ਵੇਂ' part of the input word is converted to 'ਮੇਂ'. Some rules convert the POS tags of the source text. In the following example the system is converting the 'ਰਹਿੰਦਾ' tag to 'ਰੈਂਦਾ' tag. For example, 'ਰਹਿੰਦਾ' word is converted to 'ਰੈਂਦਾ' word.

## VI. PROPOSED SYSTEM ARCHITECTURE

In this section we describe our system architecture which is shown in Fig. 3. There are three main components of the system namely Morphological Analyzer, Conversion engine and Generator. First, analyzer uses different methods to identify standard Punjabi word in a source sentence. Translation engine contains transfer rules and SP-DP corpus that perform the conversion task. The translation quality of system depends upon the efficiency of transfer rules and size of corpus. The final outcome is produced by generator component in desired dialect. The input of the system is encoded in Unicode and un-tokenized. The output of system is also encoded in Unicode.

A. *Morphological Analyzer*

The analyzer will segment the input text into single words. The main focus of the component is to identify the words of the input text and to process the input text in such a way, that it matches the data which the system is trained on. This component will examine the content of the source text to distinguish the words that will be converted to dialect Punjabi text. This system component decides which words to translate and which words to leave.

B. *Conversion Engine*

Conversion engine is the main component of proposed system. Various SP source text words identified by morphological analyzer are converted to its DP corresponding words. The training is done in two steps. In the first step, the selected words of SP are mapped with its equivalent DP words using bilingual dictionaries. The words left after first step are converted using various morphological transfer rules.

C. *Word to word mapping*

In this step, 2 bilingual dictionaries are used for direct word to word mapping. The bilingual dictionaries contain words of *Malwai* dialect and *Doabi* dialect. For example 'ਅੱਗੇ' word is mapped to 'ਮੂਹਰੇ' word. In the example direct word to word mapping is done as the whole word 'ਅੱਗੇ' word is mapped to 'ਮੂਹਰੇ' word.

D. *Using Morphological Transfer rules*

Transfer rules are used for conversion of the words left after first step. These rules are used to replace specific portion of SP source text. For example, 'ਵਰ' word is converted to 'ਬਰ' word. In the above example 'ਵ' part of the input word is converted to 'ਬ'.

Volume 3 Issue VI, June 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

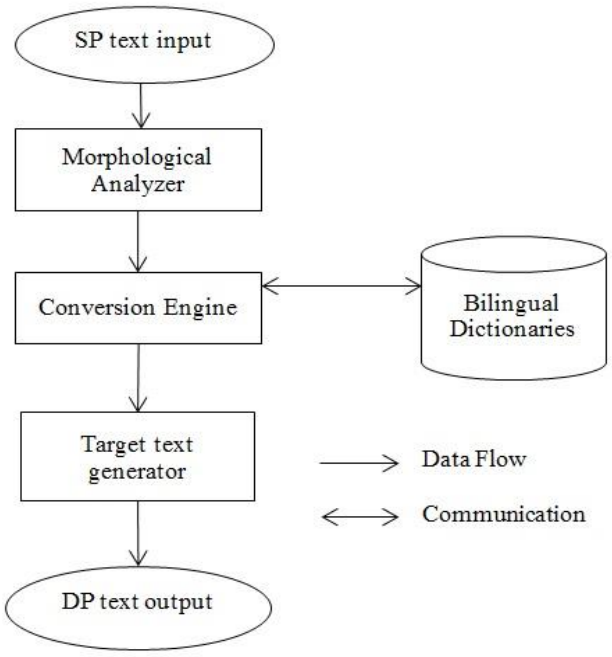



Fig. 3 Flow diagram of components of dialect conversion system

E. *Generator*

The generator component uses the previous analysis to generate different DP language words. The component replaces each unit mapped above with the corresponding unit of the desired dialect. After that, the component produces the final outcome in the desired dialect. The outcome accuracy highly depends upon the quality of translation rules and size of the bilingual dictionaries.

## VII. EVALUATION

For the evaluation purpose text collected from different literature resources have been used. A manual error analysis is conducted comparing output with same training data. The developed system is tested over 850 sentences. Out of 850 sentences, 30 sentences produce wrong conversion. Wrong conversion occurred due to spelling errors, proper nouns and mismatching of transfer rules. The results are quite promising and show the success of first dialect conversion system for Punjabi language. The current implementation can handle word order changes between SP and DP. The system performance is further improved by adding more training data.

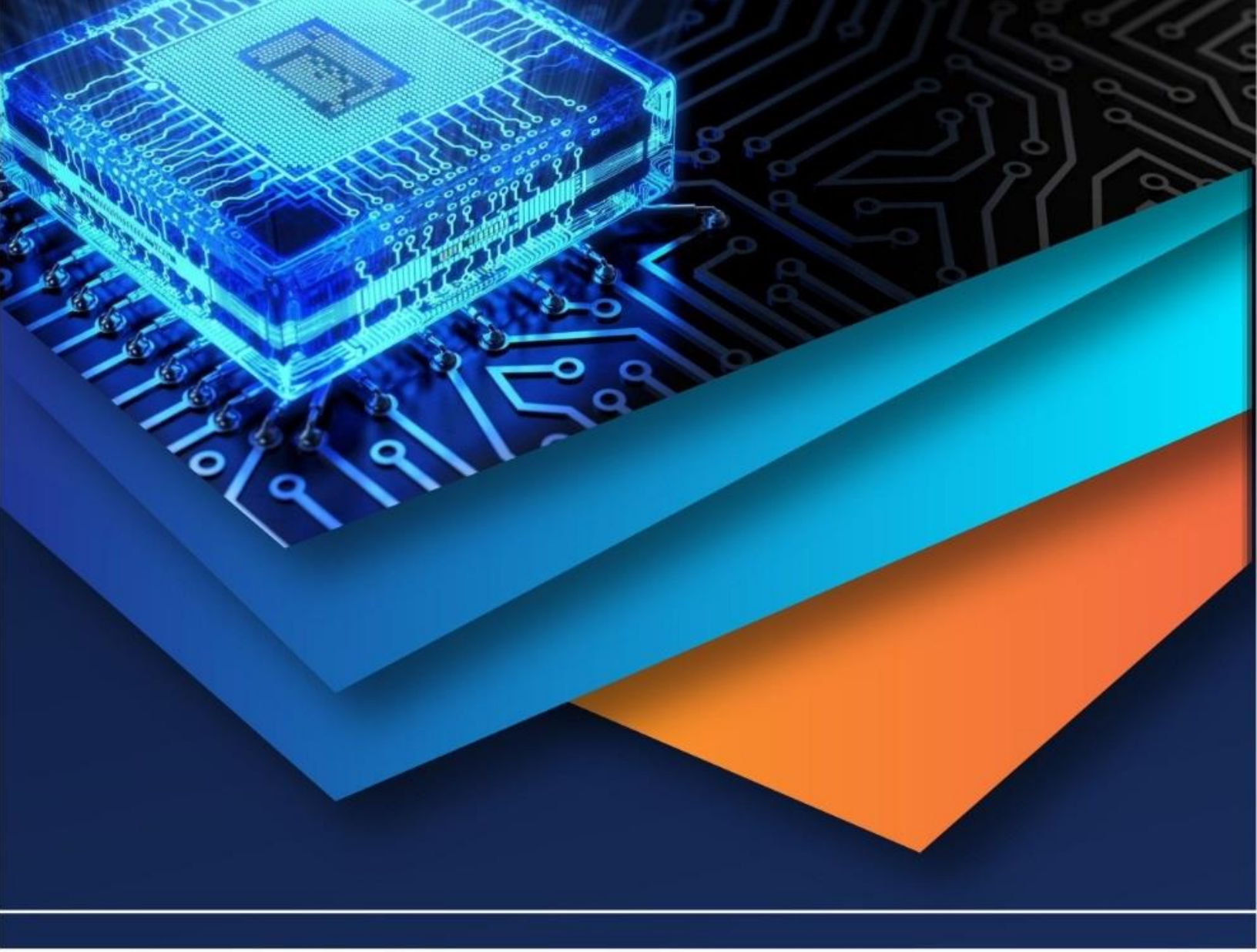
## VIII. CONCLUSION AND FUTURE SCOPE

A rule based system for converting Punjabi text from one dialect to another has been presented. There is lack of resources and processing tools for Dialectal Punjabi language. The system identifies the words from the input text, segment these into individual word and translate them to Dialectal Punjabi. The conversion engine is the main component of MT process that relies on bilingual dictionaries and morphological transfer rules. Better results are achieved by increasing the size of the training corpus. This research could help to increase the Dialectal Punjabi resources and processing tools. This is the first conversion work concerning the *Malwai* dialect and *Doabi* dialect.

In the future, the coverage of system in the handled dialects and to new dialects can be extended. The work can be done to achieve the automatic conversion of the dialect. The system can further be improved to automatically learn new morphological transfer rules from limited available data. We are interested in studying how our approach can be combined with solutions that simply add more dialectal training data.

**REFERENCES**

[1] A. John, "Two Dialects One Region: A Sociolinguistic Approach to Dialects as Identity Markers", Thesis, Ball State University, Muncie, Indiana, 2009.
[2] D. Chiang, M. Diab, N. Habash, O. Rambow and S. Shareef, "Parsing Arabic Dialects", Proc. European Chapter of ACL (EACL), 2006.
[3] H. Kaur and V. Laxmi, "A survey of machine translation approaches", International Journal of science, engineering and technology research (IJSETR), Volume 2, Issue 3, 2013, pp.716-719.
[4] H. M. A. Bakr, K. Shaalan and I. Ziedan, "A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic", Proc. 6th International Conference on Informatics and Systems, INFOS2008, Cairo University, 2008.
[5] H. Singh, "A comparative study of Doabi and Malwai dialects", M.Phil. Thesis, Punjabi University, Patiala, 2007.
[6] H. Sawaf, "Arabic dialect handling in hybrid machine translation", Proc. Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado, 2010.
[7] J. Singh, "Malwae te Poadi up-bhashvan di duni-viont", M.Phil. Thesis, Language Department, Guru Nanak Dev University regional campus, Jalandhar, 2005.
[8] K. Duh and K. Kirchhoff, "POS tagging of dialectal Arabic: a minimally supervised approach", Proc. ACL Workshop on Computational Approaches to Semitic Languages, Semitic '05, Ann Arbor, Michigan, 2005, pp. 55-62.
[9] K. Marimuthu and S. L. Devi, "Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil Text using FSTs", Proc. 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA, 2014, pp.37–45.
[10] M. Kaur, "Malwai up-bhasha da sarthik duni vigyanik adiyan", M.Phil. Thesis, Punjabi University, Patiala, 1993.
[11] M. Gillani and M. A. Mahmood, "Punjabi: A Tolerated Language Young generations'attitude", ISSN (Paper) 2224-5766, Volume 4, Issue 5, 2014, pp.129-137.
[12] N. Habash and O. Rambow, "MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects", Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006, pp.681–688.
[13] R. Kumari, "Dawabi da up-bhashai sarvekhan hoshiapur zile de parsang wich", Ph.D. Thesis, Punjab University, Chandigarh, 2002.
[14] R. Tachicart and K. Bouzoubaa, "A hybrid approach to translate Moroccan Arabic dialect", Proc. 9th International Conference on Intelligent Systems, (SITA'14), Rabat, Morocco, 2014.
[15] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. F. Zaidan, and C. C. Burch, "Machine translation of Arabic dialects", In HLT-NAACL, 2012, pp.49–59.
[16] S. Tripathi and J. K. Sarkhel, "Approaches to machine translation", Annuals of Library and Information Studies, Volume 57, 2010, pp.388-393.
[17] W. Salloum and N. Habash, "Dialectal to Standard Arabic Paraphrasing to Improve Arabic English Statistical Machine Translation", Proc. First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Edinburgh, Scotland, 2011, pp.10–21.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY