ISSN 22773061

Direct Machine Translation System from Punjabi to Hindi for Newspapers headlines Domain

[1]Sumita Rani, [2]Dr. Vijay Luxmi

[1]Student, Dept. of C.S.E, Guru Kashi University, Talwandi Sabo (Bathinda)

smita.kamboj.43@gmail.com

[2]Associate Prof., Dept. of C.S.E, Guru Kashi University, Talwandi Sabo (Bathinda)

cse_vijay2003@yahoo.co.in

ABSTRACT

Machine Translation System is an important area in Natural Language Processing. The Direct MT system is based upon the utilization of syntactic and vocabulary similarities between more or few related natural languages. The relation between two or more languages is based upon their common parent language. The similarity between Punjabi and Hindi languages is due to their parent language Sanskrit. Punjabi and Hindi are closely related languages with lots of similarities in syntax and vocabulary. In the present paper, Direct Machine Translation System from Punjabi to Hindi has been developed and its output is evaluated in order to get the suitability of the system.

Index Terms: Machine Translation, Direct Approach, Word Sense Disambiguation, Transliteration, Punjabi, Hindi.

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 8, No 3

editor@cirworld.com
www.cirworld.com, member.cirworld.com

I. INTRODUCTION

Machine Translation system is software designed that basically takes a text in one language and translates it into another language. There are number of approaches for Machine Translation System like Direct approach, Transform based approach, Interlingua based approach, Statistical approach, Rule based approach, web based Approach etc but the choice of approach depends upon the kind of languages involved. If two languages are structurally and vocabulary similar then direct approach is the best choice for the development of Machine Translation System for these related languages. Although Punjabi and Hindi are in the same family, but still they have many difference in order to make them not mutually Intelligible. Mutual intelligibility of the languages depends on the degree of phonetical, morphological, syntactical and lexical similarity. Both languages are not mutually intelligible in written form but in spoken form, both are mutually intelligible.

II. SYSTEM DESCRIPTION

i. Synthetic vs Analytic.

The parent language (Sanskrit) of Punjabi and Hindi is Synthetic. Its descendent Hindi is Analytic. It means we need to add some words, known as preposition or postpositions, to convey the relation in Hindi. E.g. in Punjabi we say

Punjabi: ਕਾਰ ’ਚ ਮੋਬਾਈਲ ਵਰਤਣ ’ਤੇ ਸਾਰਾ ਗ੍ਰਿਫਤਾਰ

And in Hindi the same meaning is conveyed by adding preposition as

Hindi: कार में मोबाईल के उपयोग पर सारा गृफतार

Such type of pre position is must in Hindi. It shows the relationships among subject and object. In comparison to Hindi, Punjabi is also Analytic but not completely. There are many examples which show that Punjabi is synthetic also. As for example consider the sentence below:

Punjabi: ਦੋ ਨੌਜਵਾਨਾਂ ਦੀ ਡੁੱਬਣ ਕਾਰਨ ਮੌਤ

Hindi: दो युवकों की डूबने कारन मृत्यु

Although it is intelligible by the Hindi person who reads but grammatically it is not correct. The above sentence should be translated as follow:

Hindi: दो युवकों की डूबने के कारन मृत्यु

Few rules can be employed to take care of these types of problems.

ii. Ambiguity

The ambiguity problem is one of the major problems in any machine translation system. Two types of ambiguity problems are presented here i.e. structural ambiguity and word level ambiguity.

There is no structural ambiguity in the Punjabi language that does not carry over as such in Hindi language because Punjabi and Hindi both are structurally same. But in some sentences this problem is evaluated.

Word level ambiguity is a problem in translating Punjabi to Hindi language.

Punjabi: ਭਾਅ ਜੀ ! ਮੈਨੂੰ ਥੱਲੇ ਲਾਹੋ

Hindi: भैया जी ! मुझे नीचे उतारो

Punjabi: ਰੂੰ ਦੇ ਭਾਅ ਡਿੱਗਣ ਨਾਲ ਤੇਜੜੀਆਂ ’ਚ ਹਲਚਲ

Hindi: रुई के भाव गिरने से तेजड़ीयों में हलचल

Here the one Punjabi word **ਭਾਅ** has two meanings in Hindi one corresponds to **भैया** and another corresponds to **भाव**.

Direct translation does not provide any solution for these types of problems. Another methodology or rules are developed for such types of problems.

iii. Gender conversion

Some words are changed in target language due to their gender conversion.

In Punjabi: ਦੀ ਵਰਤੋਂ

In Hindi: का उपयोग

In Punjabi: ਦੀ ਜ਼ਰੂਰਤ

In Hindi: की ज़रूरत

In the above sentences, the Punjabi word **ਦੀ** is translated in Hindi language according to their gender representation. The word **ਵਰਤੋਂ** is in the male sense and the word **ਜ਼ਰੂਰਤ** is in the female sense.

iv. Double representation of plural form of nouns

In Punjabi language almost all plural forms of nouns of the Punjabi language is double representation in Hindi language. E.g. **ਔਰਤਾਂ** can be converted into **औरतों** or **औरतें** depending on the presence of position.

Punjabi: ਔਰਤਾਂ ਦੀ ਸੁਰੱਖਿਆ ਲਈ ਹੋਰ ਕਦਮ ਚੁੱਕਣੇ ਪੈਣਗੇ

Hindi: औरतों की सुरक्षा के लिए और कदम उठाने पड़ेंगे

Punjabi: ਹੁਣ ਵਧੇਰੇ ਜ਼ਿੰਮੇਵਾਰੀ ਵਾਲੀ ਨੌਕਰੀ ਤੋਂ ਨਹੀ ਝਿਜਕਦੀਆਂ ਔਰਤਾਂ

Hindi: अब अधिक जिम्मेवारी वाली नौकरी से नहीं कतराती औरतें

III. SYSTEM ARCHITECTURE

The resulting system architecture is shown in the following stages through which the source text is passed.

i. Preprocessing Stage

Collections of operations are applied on input data to make it process able by the translation engine in the preprocessing stage. In our current working system, we have performed following pre processing steps:

a. Text normalization

There are number of Unicode fonts to represent Punjabi text and each font has variations in assigning Unicode code to Punjabi Alphabets so the first step is to normalize the source text by converted it into Unicode format. If the text is in Unicode format then it works on any computer system without installing the Punjabi typing font. The output is in Unicode format can be used in various ways in various applications.

b. Tokenization

Our system is designed to do sentence level translation. Once the whole text is scanned, next step is to break up the data into sentences then individual words or tokens (characters) are extracted out from the sentence and its equivalent token in the target language is found out. The process is repeated for all the words of input data.

ii. Translation Engine

The translation engine is responsible for translation of each token obtained from the previous step. It uses various lexical resources for finding the match of a given token in target language. The following steps are used in this phase.

1. The token is checked for proper word then the word is matches with in the data base if the similar word is present in the data base then it translate from source language to target language.
2. If the matching word is not presented in the database then comparing the word and check the preposition or postposition of the word and then apply the rule according to its current position.

iii. Transliteration Engine

Some words like name, sir name, city name, country name, idioms etc does not need to be translated because such types of words has same pronunciation in both Punjabi and Hindi language. These types of words are transliterated by direct character to character mapping.

IV. TESTS AND EXPERIMENTS

	General News	Politics News	Business News	Sports News	Entertainment News	Other News	Total
Total News Headlines	100	100	100	100	100	100	600
Total Words	1047	1207	1081	1082	965	804	6186

i. For intelligibility test

It is a subjective test which is used to check the intelligibility of the system. Intelligibility is effected by grammatical errors, miss-translations, and un-translated words. A Grade scale point is made in which highest grade point is assigned to those sentences that look perfectly accurate and understandable and the lower grade point is assigned to the sentence which is understandable. This grade point scale look likes:

A	Completely Faithful
B	Fairly faithful: more than 50 % of the original information passes in the translation.
C	Barely faithful: less than 50 % of the original information passes in the translation.
D	Completely Unfaithful. Doesn't make sense.

Grade scale point according to sentences is as follows:

Grade point	Punjabi	Hindi
A	ਹਰਿਆਣਾ ਤੇ ਰੇਲਵੇ ਵਿਚਕਾਰ ਹੋਵੇਗਾ ਫਾਈਨਲ	हरियाणा और रेलवे के बीच होगा फाईनल
B	ਭਾਰਤ ਨੇ ਚੀਨ ਅਤੇ ਬ੍ਰਾਜ਼ੀਲ ਦੇ ਮੁਕਾਬਲੇ ਜ਼ਿਆਦਾ ਵਾਧਾ ਦਰਜ ਕੀਤਾ	भारत ने चीन और ब्राज़ील के प्रतियोगिता ज़्यादा वृद्धि दर्ज किया
C	ਮੈਂ ਵੀ ਵਿਆਹਿਆ ਹੋਇਆ ਹਾਂ	मैं भी शादीशुदा हुआ हूँ
D	No sentence found	

ii. For Accuracy test

A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called accuracy. For the accuracy test, those sentences are summed up which are properly translated and not properly translated and which are badly or not translated. Percentage of accurate translation system is calculated by counting down the number of accurate sentences.

V. RESULTS

Average ratings of %age for the system is depend upon the intelligible test and accuracy test. Accuracy is measured with the help of four grade scale points. From the accuracy analysis total number of accurate sentence are calculated and then their %age is found out which is come out to be 97%.Error rating of the system depends upon the wrongly translated word or expression, Un-translated words and Wrong choice of words. The %age of error rating is found out from the total sentences which is come out to be 3%.

VI. CONCLUSION

The accuracy of the translation achieved by our system justifies the suggestion that word-for-word translation for machine translation system for language pair of Punjabi-Hindi. The major inaccuracies in the direct translation are due to poor word choice for confusing words and some corrections regarding preposition and post positions. The lack of information about

Hindi language is sometimes causes an unnecessary translation error. We can conclude that this study is beneficial to remove the language barrier between two closely related language pair Punjabi-Hindi.

VII. REFERENCE

[1] Gurpreet Singh Josan and Gurpreet Singh Lehal, "Direct Approach for Machine Translation from Punjabi to Hindi" CSI Journal of Computing | Vol. 1, No.1, 2012.

[2] Gurpreet Singh Josan1& Jagroop Kaur, "Punjabi to Hindi statistical machine transliteration" International Journal of Information Technology and Knowledge Management July-December 2011, Volume 4, No. 2, pp. 459-463.

[3]. Vishal Goyal and Gurpreet Singh Lehal, "Hindi to Punjabi machine translation system" Proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6, Portland, Oregon, USA, 21 June 2011. Association for Computational Linguistics.

[4]. Vishal Goyal and Gurpreet Singh Lehal, "Web based Hindi to Punjabi machine translation system" journal of emerging technologies in web intelligence, vol. 2, no. 2, may 2010.

[5] Gurpreet Singh Josan and Gurpreet Singh Lehal, "A Punjabi to Hindi Machine Transliteration System" Computational Linguistics and Chinese Language Processing Vol. 15, No. 2, June 2010, pp. 77-102 The Association for Computational Linguistics and Chinese Language Processing.

[6] Vishal Goyal and Gurpreet Singh Lehal

"Evaluation of Hindi to Punjabi Machine Translation System" IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009.

[7]. Gurpreet Singh Josan and Gurpreet Singh Lehal, "Evaluation of Direct Machine Translation System For Punjabi To Hindi"

[8] Vishal Goyal and Gurpreet Singh Lehal, "Hindi-Punjabi Machine Transliteration System (For Machine Translation System)"

[9] Online Punjabi news headlines from jagbani newspaper http://www.jagbani.com/news

[10] Daily Online news headlines http://beta.ajitjalandhar.com/news/

[11] Punjabi news from http://punjabi.samachar.com/

Author's biography

Ms. Sumita Rani received her B-Tech degree in computer Engineering from Yadavindra College of Engineering, Talwandi Sabo (Bathinda) in 2011 and M-Tech degree in computer science & Engineering from Guru Kashi University, Talwandi Sabo (Bathinda). Her Research area is Natural Language Processing and Machine Translation.

Dr Vijay Laxmi received her B-Tech degree in computer science & Engineering from SLIET Longowal in 2003, and Ph.D degree in Computer Science & Engineering in year 2012. Her research area is Grid Computing & OCR. She has published 25 research papers in various National/International Conferences and Journals. At present, she is engaged in Guru Kashi University, Talwandi Sabo, and Punjab as Dy. Dear Research and an Associate Professor in Computer Science Engineering & Information Technology department.