Issues in Conversational Sanskrit to Bhojpuri MT

Shagun Sinha, Girish Nath Jha
Jawaharlal Nehru University, New Delhi
{shagunsinha5, girishjha} @gmail.com

Abstract

The authors of this paper have presented an alpha version of MT system for conversational Sanskrit to Bhojpuri. The paper discusses the challenges in corpora creation for less resourced languages of India, training and evaluation of the MT system in the domain of everyday conversation and the research questions emerging from there.

Keywords: Statistical Machine Translation, Less Resourced Languages Technology, Sanskrit to Bhojpuri.

1.Introduction

With over 30 million speakers as per the 2001 census data, Bhojpuri (ISO : 630-9.[1]) is an Indian language spoken primarily in the state of Bihar in the districts of Champaran, Saran, Shahabad districts; Assam, parts of Uttar Pradesh and West Bengal. [2] It is not listed in the eighth schedule of the Indian Constitution. Bhojpuri is a "morphologically rich and non-configurational language, unlike English" (Behera et al, 2016).

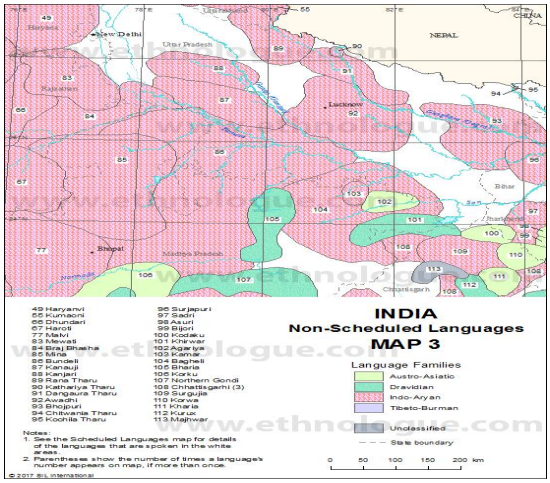

Fig 1. Map of the Non-scheduled languages of India Source: https://www.ethnologue.com/map/IN_03

Bhojpuri is a less resourced language and has found very little space in the technological space. Singh (2014) prepared a POS tagger for Bhojpuri as per the BIS scheme which is the first of its kind work (Singh and Banerjee, 2014). Singh and Jha (2015) proposed a statistical tagger for Bhojpuri based on Support Vector Machine (SVM) whose results were 87.67% for test and 92.79% for gold corpus. A paper based on 'Bhojpuri Annotation' was presented by Neha Mourya in 2015.[3] She also authored papers on "Complex Predicate in Bhojpuri," "Classification of Bhojpuri Adverbials," "Agreement in Bhojpuri"(ibid). Behera et al (2016) worked on "Dealing with Linguistic Divergence in English-Bhojpuri Machine Translation" in which they use Dorr's theoretical framework for resolving divergences between the two languages. Linguistic Divergence "...refers to the concept of structural or 'parametric variation' between a source language (SL) and a target (TL) pair in Machine Translation(MT). In other words, it emerges when the decoded output content lacks 'well-formedness' because of the inherent` linguistic constraints" (Behera et al 1, last accessed Jan 14, 2018). Some research in area of MT is currently in progress at the Jawaharlal Nehru University, New Delhi (JNU) the results for which are currently awaited.

Sanskrit, on the other hand, has been a language of the sub-continent for approximately more than 6000 years. The literary traditions so developed in this 'donor' language are a prized heritage. Many Indian languages have drawn from it from time to time to strengthen their literature. Some MT tools for Sanskrit have been developed with varying results. Pandey (2016) has developed a Sanskrit to Hindi MT as a part of his Ph.D. work at JNU.

In developing resources for Bhojpuri, Sanskrit would prove to be a good literary resource. Additionally, Sanskrit can be taught to Bhojpuri speakers too. To teach Sanskrit to Bhojpuri community, e-learning or other learning technologies can be important. A conversational text would assist in preparing tools for teaching Sanskrit to Bhojpuri speakers.

The current work is a first of its kind at initiating an MT for Sanskrit to Bhojpuri based on a corpora of everyday conversational text.

---

[1] Ethnologue: https://www.ethnologue.com/language/bho

[2] ibid

[3] https://fmsbhu.academia.edu/NehaMaurya/CurriculumVitae accessed 14th Jan 2018

2. Language Technology Resource (LTR) Creation for Bhojpuri and Linguistic Analysis

The entire set of parallel data was raw and had not been annotated for its properties. The process of parallel corpora preparation had two levels. First, collection of data, and second, translation to Bhojpuri. At the first stage, there were two further possibilities. One, the data had to be directly taken from Sanskrit sources. Second, in cases where the source was not Sanskrit but English (as in case of some news channels), the data had to be first transcribed, then be first translated into Sanskrit. At the second level, the data was finally translated into Bhojpuri for creating the parallel corpora.

2.1. Role of conversational text:

A set of conversational parallel data was one of the key sources of good translations. The parallel corpora contained short sentences of up to 5-6 words. The motive was to train an MT system with corpora that had the least of ambiguities and thus, short conversational texts were preferred. Simple conversational text was taken from news channels to avoid literary or metaphorical usage.

2.2. Corpora Creation

The corpus was created using a multitude of sources of spoken Sanskrit. The online version of Vyavaharsahastri[4] published by Samskrita Bharati (ibid) was used. Online lectures of spoken Sanskrit were transcribed to be translated later (CecUgc)[5]. Not more than 5 YouTube videos of news channels were also used to arrive at a proper conversational Sanskrit text.

The collection of sentences was based on conversational properties- the sentences had simple syntactic structures. The files so created involved naming along the ISO name codes, ie, sa for Sanskrit and bho for Bhojpuri.

In a previous research carried out on a smaller data set, the BLEU was 33.51. For advancing the system performance, the corpora size was increased to a total of 10k sentences. With not more than 7 words each, the first 5k sentences were the simplest set of sentences. The next set of 5k sentences were more complex with 8 words each on an average. After building that system, this set was further divided into different sentence sets for training different systems on even smaller data size. Next, all such separate files of smaller data size were collectively set for training taking the total number of extracted sentence count to 64,843; the number of aligned sentences to 29,669 and the number of used sentences to 59,365. The system thus trained on a corpora which had repetitions of sentences due to repetition of sets. The sentences so collected in the corpora were all utilised for training the system.

Tiwari (1960) has mentioned a detailed analysis of the language spoken in various places. Three types of Bhojpuri are spoken[6], namely, Standard Bhojpuri; Western Bhojpuri and Nagpuria. Standard Bhojpuri is spoken in THE area in and around Bhojpur (ibid).
The standard form has been explained by Tiwari in his work titled 'The Origin and Development of Bhojpuri' (1960). The forms of the male singular words used in Bhojpuri are as given below and have used standards as mentioned by Tiwari and Upadhyay (2008). The post-positions cited by Tiwari are as follows:

Case	Sa	bho
Nominative	Rama	rAma/Ramuvaa[7]
Objective	rAmaM	rAma/ ramuvA'ke' (Tiwari 111)
Instru.	rAmeNa	rAma/ramuvA se (109)
Dative	rAmAya	rAma/ramuvAKAtira
Ablative	rAmAt	rAma/rAmuvA'se' (109)
Genitive	rAmasya	rAmuvA'ke' (111)
Locative	rAme	rAmuvapa/para/ 'mein' (111)

Table1. Post-Positions in Bhojpuri (Tiwari 1960)

Pronouns are similar to standard Hindi except in second person when the use for second person singular is mostly 'tU/tohanI'[8] or rauwA for respect (ibid). For third person, the use goes to 'hama' mostly (ibid).

Verbs formed in Bhojpuri indicate usually the time and person (first, second, third) doing the act also. An example of Present Perfect as cited in Tiwari (1960) 182 is being given here:

Verb	Sanskrit	Meaning
U gayilasa	sah agachchat	He went
U gayalI	sA agachchat	She went
U gayilan	sah agachchat	He went (sense of honor)

Table 2. Verbs in Bhojpuri (Tiwari, 1960 182)

Similar has been indicated for Past perfect which is indicated using the word 'raha' (Tiwari 1960 p185),

[4] Sanskrit Ebook Website, See Ref List

[5] Learn Sanskrit Be Modern series

[6] Upadhyay 21

[7] Upadhyay, 26

[8] Upadhaya 26

Past continuous indicated by 'raha' (Tiwari 1960 p182).

3. Using MT Hub for Training

Microsoft Translator Hub is an extension of Automatic Microsoft Translator API service which has been "designed for organizations that have specific translation needs" (https://www.youtube.com/watch?v=b5qBSlKwDeg) and has been quite useful in training MT systems for research and development. The successful systems can also be deployed for wider testing evaluation by the community.

For the training part, the aforementioned corpora and parallel corpora is utilised. The parallel corpora is built by the Hub upon the upload of each of the Sanskrit monolingual corpora file and the translated Bhojpuri corpora file separately. The two files are merged to form a Parallel corpora file. Additionally, the monolingual Bhojpuri file is uploaded separately for providing monolingual reference.

The training takes place on a set of training data which it takes from the uploaded files. A section of the parallel corpora which is different from the training data is set apart for the parallel file deriving references from the monolingual or target language corpora. Upon training the system on a set of data, the hub tests it and that produces the BLEU score based on the degree of similarity with the Human reference provided in the corpora.[9]

In a previous attempt, the system had a simple set of 5k sentences for parallel data and the BLEU was 33.51. The data size, however, was low and thus, not sufficient.

4. Evaluation of the MT output

For the current research. the result obtained at the end of training was a BLEU score of 37.28. The system took 1 hour 28 minutes to train.

The sentence translations obtained in a total of 42 pages can be divided into these categories:

- Sanskrit words mixed with Bhojpuri & mixed inflections
- Exact Sanskrit Phrase used in Bhojpuri
- Clear Perfect Translations
- Clear meaning despite unmatched references
- Translation better than the reference

[9] https://www.youtube.com/watch?v=-UqDljMymMg last accessed 03.03.2018 at 23.50 hrs IST

4.1. Mixed Translations

These translations included Sanskrit words also. For example, in the following sentences with the Reference pattern given as on the Microsoft Translation platform where P stands for Page, S for Sentence number:

Reference	Sanskrit	Bhojpuri
P5, S2	eSha auShadhih	I **auShadhih** bATe
P3, S1	sImnah ullaMghanaM karoti sarvadA	sImA ke **ullaMghanaM** kare lA hamesA

Table3. Mixed Translations as obtained

These occurrences were prevalent and prominent throughout the results.

Mixed Inflections

Reference	Sanskrit	Bhojpuri
P6, S9	kaH rakShayiShyati deSam	ke **rakShayi Shyati log**

Table4. Mixed Inflections as obtained

In total, the mixed translations occurred in 2-5 out of 10 sentences.

4.2. Exact Sanskrit words:

Reference (as in the Translation platform)	Sanskrit	Bhojpuri
P24, S10	gRhavyavasthA	**gRhavyavas thA**

Table 5. Exact Sanskrit Words

Such translation happened **only in single word** sentences which did not have repeated occurrences.

4.3. Perfect Translations

Reference (as in the Translation platform)	Sanskrit	Bhojpuri
P29, S1	kasmAt sarve bibhyati	kekarA se saba Dare lana
P29, S3	kim abhavat	kA bhayila

Table 6. Perfect Translations as obtained in results

Perfect translations ranged from 0-3 occurrences for every 10 sentences.

4.4. Correct meaning despite unmatched reference

There were instances where the meaning was conveyed despite being unmatched to the reference. It occurred in two key forms one of which was change in morpheme order:

Reference (as in the Translation platform)	Reference	MT
P33, S2	aba kaa karIM hamanIM	aba hamanI kA karIM
P21, S1	phera trikoNa **paDhIM lA**	phera trikoNa **paDhaba**

Table 7.Correct Meaning for unmatched reference.

An average of such occurrences after considering 10 result pages indicates that such instances have an average occurrence of 38%, i.e., 38 times every 100 sentences.

Another observation in this category includes the presence of words different from the reference but synonymous with it. For example, reference sentence 40/11 : The MT uses the word 'laIkAI' for the reference word 'bachpan' both of which mean childhood or youth. Another example includes the translation of 'ehi' instead of 'IhI' both of which indicate the same meaning. The occurrence of such instances, however, is in not more than 3 out of 11 sentences approximately.

4.5. Translation better than the Reference

An improvement from the previous attempt was indicated in instances of the MT using translated words which were better than those given in the reference. For example, On P41 S10, the word for 'priya' (favorite) in Bhojpuri was translated as **pasandIdA** which is a more local version of the standard reference 'priya'. Other instances include '**bahut sArA'** instead of '**kayiyan' (many), 'bujhAyila'** instead of '**janalas' (came to know).** Such occurrences were between 0-2 times every 20 sentences.

Indication

Four indications can be obtained from the results. First, short sentences are easily and well translated.

Second, the words where inflections were retained were, in many cases, the instances where the word had a closely similar word in Bhojpuri also.

Third, the pattern of verbs was successfully translated at every instance. 'cAha tA' for Sanskrit 'ichchhati' (wants), 'hova tA' for 'bhavati' Sanskrit for 'is happening' were translated with good degree of accuracy.

Fourth, there are no instances of zero matched references. This means that the training succeeded to a great extent. At least one word was definitely translated correctly in the test results obtained after training.

The results of occurrence (every 10 sentences) can be summarized as in the table below:

S.No.	Type	Occurrence
1.	Mixed	2-5
2.	Exact Sanskrit words	≤1
3.	Perfect Translation	0-3
4.	Correct Meaning despite unmatched reference	3.8
5.	Translation better than reference	≤1

Table 8. Overall Evaluation of the Results

5.Conclusion

The current work is first of its kind work in the area MT for Sanskrit to Bhojpuri using conversational corpus by the same authors.

Use of conversational corpora enabled the presence of short sentences . This enabled easy manual translation as well as better alignment.

Similarly, the rise in BLEU score indicated that repetition of reference sentences to the system actually aids in better training.

On such a use, mixed translations which conveyed the overall meaning well emerged as the highest occurring results. Training with better set of data will be able to conclude in higher number of perfect translations.

Further work must be initiated in the direction with more data and larger corpora. Use of simple split-prose text for creation of additional parallel aligned data would enhance the results . Additionally, literary sources with metaphorical use of language may be included in the training data for wider coverage.

Acknowledgements

The authors of this work based their work on the MTHub platform of Microsoft. They are thankful for having been given the platform. Also, Dr. Sankara, Dr. Himanshu Pota and Prof. Baldevanand Sagar, and Bhagini Manjushree helped immensely in the

collection of Sanskrit sentences for which the authors extend their deepest gratitude to them.

References

Behera Pitambar, Neha Maurya and Vandana Pandey. "Dealing with Linguistic Divergences in English-Bhojpuri MT". In: *Proceedings of the South and Southeast Asian Natural Language Processing.* 2016: Osaka, Japan. Pages 103-113. Accessed on: ResearchGate

Maurya, Neha. "Complex Predicate in Bhojpuri". Presented at the 33rd AICL organized by Punjab University, Patiala.

_____. "Classification of Bhojpuri Adverbials" Presented at 9th ICOSAL organized by Punjab University, Patiala.

_____. "Agreement in Bhojpuri." Presented at 11th ICOSAL at BHU, Varanasi.

_____. "Bhojpuri Annotation." Presented at regICON organized by IIT-BHU: Varanasi, 2015.

Pandey, Rajneesh. *Sanskrit-Hindi Statistical Machine Translation: Perspectives and Problems.* Ph.D. Thesis: JNU New Delhi, 2015.

Singh, Srishti. *Challenges in Automatic POS Tagging of Indian Languages: A Comparison of Hindi and Bhojpuri POS Tagger.* M.Phil. Thesis: JNU New Delhi, 2015.

_____ and Esha Banerjee. "Annotating Bhojpuri Corpora using BIS scheme." In: *Proceedings of the WILDRE Conference,* Iceland: 2014.

Sinha, Shagun. *Translation Issues in Conversational Sanskrit-Bhojpuri Language Pair: An MT perspective.* M.Phil. Thesis (Unpublished) : JNU New Delhi, 2016.

Tiwari, Uday Narayan. *The Origin and Development of Bhojpuri Language.* Kolkata: The Asiatic Society, 1960. 2001 Reprint.

Upadhyay, Krishandev. *Bhojpuri Loksahitya.* Varanasi: Vishwavidyalaya Prakashan, 2008.

Vyavaharasahastri. Samskrita Bharati.

Ethnologue website: https://www.ethnologue.com/map/IN_03

Vyavaharsahasri link: http://www.sanskritebooks.org/2009/04/sanskrit-daily-conversation/

YouTube Videos: https://www.youtube.com/watch?v=HTrDZ_5YXow

https://www.youtube.com/channel/UCwqusr8YDwM-3mEYTDeJHzw

https://www.youtube.com/user/ndtv

https://www.youtube.com/watch?v=-UqDljMymMg