

Incorporating Localised Context in Wordnet for Indic Languages

S. H. Mohapatra¹, S. Agnihotri¹, A. Garg, P. P. Shah, S. Chakraverty

Netaji Subhas University of Technology

Dwarka, New Delhi

{soumyam, shikhara, aparg, praveens}.co.16@nsit.net.in, shampa@nsit.ac.in

Abstract

Due to rapid urbanization and a homogenized medium of instruction imposed in educational institutions, we have lost much of the golden literary offerings of the diverse languages and dialects that India once possessed. There is an urgent need to mitigate the paucity of online linguistic resources for several Hindi dialects. Given the corpus of a dialect, our system integrates the vocabulary of the dialect to the synsets of IndoWordnet along with their corresponding meta-data. Furthermore, we propose a systematic method for generating exemplary sentences for each newly integrated dialect word. The vocabulary thus integrated follows the schema of the wordnet and generates exemplary sentences to illustrate the meaning and usage of the word. We illustrate our methodology with the integration of words in the Awadhi dialect to the Hindi IndoWordnet to achieve an enrichment of 11.68 % to the existing Hindi synsets. The BLEU metric for evaluating the quality of sentences yielded a 75th percentile score of 0.6351.

Keywords: IndoWordnet, geographical wordnets, lexicons, clustering

1. Introduction

The Hindi Belt or Hindi heartland, is a linguistic region consisting of parts of India where Hindi and its various dialects are spoken widely (Sukhwil, 1985). Hindi as a language has evolved over the years due to migration and invasion of various socio-ethnic groups like Turks, Britishers etc. This has given it a dynamic shape with the Devanagari script remaining nearly the same but the speech changing with location, thereby leading to a plethora of dialects spread across the region. Of late, due to westernisation and globalization, over 220 Indian languages have been lost in the last 50 years, with a further 197 languages marked as endangered according to People's Linguistic Survey of India, '18 (V. Gandhi, 2018). There is thus an exigent need to preserve not only the Hindi language but its various dialects that give India its unique identity embodying unity in diversity. Through this project we take a step towards protecting this linguistic heritage.

The Indo-wordnet is a linked structure of wordnets of major Indian languages from the Indo-Aryan, Dravidian and Sino-Tibetan families. It was created by following the expansion approach from Hindi wordnet which was made available free for research in 2006 (Bhattacharya, 2006). However, each Indic language has a number of dialects for which the IndoWordnet has no related information. In this paper, we enhance this digital footprint by systematically incorporating vocabulary from Hindi dialects using language processing tools, algorithms and methods.

Our research contributes by first collating the resources of a dialect that are available in Devanagari script from multiple textual sources as well as from audio clips of real conversations. This consolidated corpus is subsequently used to extract the vocabulary and exemplify

its appropriate usage to enrich the indowordnet. Ultimately, the wordnet is envisioned to be a complex whole containing not just word usages but the way a particular word is pronounced and used in different geographical regions. We demonstrate our methodology by using the Awadhi dialect to enrich the Hindi IndoWordnet.

2. Prior Work

Of late, several research groups have contributed towards enrichment of wordnet for different languages.

Researchers from Jadavpur University (Ritesh, 2018) have developed an automatic language identification system for 5 closely-related Indo-Aryan languages of India namely, Awadhi, Bhojpur, Braj, Hindi and Magadhi. They have compiled corpora with comparable format but varying lengths for these languages by tapping upon various resources.

Mikhail et al. (Mikhail, 2017) present an unsupervised method for automatic construction of WordNets based on distributional representations of sentences and word-senses using readily available machine translation tools. Their approach requires very few linguistic resources and can thus be extended to multiple target languages.

Nasrin Taghizadeh et al. (Nazrin, 2016) propose a method to develop a wordnet by only using a bi-lingual dictionary and a mono-lingual corpus. The proposed method has been executed with Persian language. The induced wordnet has a precision of 90% and a recall of 35%.

Taking inspiration from the above approaches, we formulate our own approach and propose an algorithm to build wordnets for low resource dialects of Hindi. Our work primarily focuses on dialects, an area which has thus far been ignored.

¹ These authors have contributed equally

3. Method

We present an algorithm that takes in the corpus of a given dialect and its Hindi bilingual dictionary. The system uses this dialect's vocabulary to enhance the synsets of Hindi IndoWordnet. Our twin goals are to ensure that the enriched vocabulary follows the schema of the wordnet and that we are able to generate exemplary sentences to illustrate the meaning and usage for each dialect word.

3.1 Data Collection

We used the Awadhi-Hindi bilingual dictionary, also called *Awadhi Shabdkosh*², an ebook, for extracting Awadhi words along with their relevant Hindi meanings. We used this collection solely for the purpose of IndoWordnet synset integration.

The main source of our corpus compilation came from the comparable corpora from Jadavpur University (Ritesh, 2018) which consists of a training set of 70350 lines, a validation set of 10300 lines, test data of 9600 lines and 9600 lines of gold data for test sentences. The gold data contains the labels for the test data. The comparable dataset consists of tagged sentences belonging to Awadhi, Bhojpuri, Braj, Hindi and Magadhi categories.

A total of 12297 Awadhi sentences were extracted from the Jadavpur corpus on the basis of these tags. We also found other sources of Awadhi literature in electronic form including an ebook containing Bible Stories in Awadhi and audio samples of conversations on social topics³. These additional resources yielded 3500 Awadhi sentences. The consolidated dataset is used for creating an Awadhi lexicon of preferential pairs of Awadhi words and for sentence generation.

3.2 Automatic Mapping of Existing Resources

IndoWordnet provides the NLP resources of various Indic languages. However, it does not store any linguistic information about the various dialects in which a word may be spoken in different regions. In this section, we explain the processes involved in mapping the Awadhi words in the bilingual dictionary to the relevant Hindi synsets.

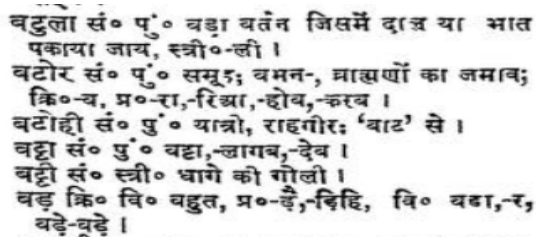


Figure 1: Image from the *Awadhi Shabdkosh*: The Awadhi-Hindi bilingual dictionary

3.2.1 Preparation of Inverse Dialect Mappings

An excerpt from the *Awadhi Shabdkosh* is shown in Figure 1. An awadhi word (say, बटोर) is followed by its POSTag (सं, for noun), its gender (पु for male) and its meaning in hindi. The symbol '।', known as the purnaviram, marks the end of a sentence.

Using an Optical Character Reader⁴ (OCR) and regex, we bring the PDF data into usable textual format.

We pick each Indic word from the inverted mappings and search it in the IndoWordnet. If it exists then we integrate the dialect word into the first synset of the Indic language and add a metadata tag to the same indicating the dialect to which it belongs. The regex used to generate inverted mappings is shown in Figure 2.

```
Regex = r“( [०पु०] [\u0900-\u0965\u0967-\u097F]+ (,|:|;| | [\u0900-\u0965\u0967-\u097F]+) )”g
```

Figure 2: Regex used to extract mappings from plain text

The regex helps us to extract pairs of Awadhi-Hindi words with each Awadhi word having a single word Hindi meaning so as to preclude superfluous words. Each regex match has 3 groups. The green group picks up the POSTag delimiter, the blue group picks up the single Hindi word we are concerned with and the red group picks up the delimiter for the Hindi word. The Hindi word delimiter group helps filter out single word meanings.

Figure 3, shows different matches detected using the regex 101 tool⁵. To illustrate, in the first match, the Awadhi word is लड़कपिल्ली, the POSTag delimiter is ०, the Hindi word is लड़खड़ाब and Hindi word delimiter is किर०. Hence, the Awadhi-Hindi pair is लड़कपिल्ली - लड़खड़ाब.

3.2.2 Using Metadata Tags

With the integration of dialect related lingual information for each Indic language in the IndoWordnet, it becomes mandatory to add metadata tags to each word to indicate its membership in different dialects. Adding metadata tags enriched the semantic information of each word in the IndoWordnet.

For every new word that is being integrated to the IndoWordnet, the metadata tag denotes the dialect from which the request for integration into the wordnet has been instantiated. Figure 4a shows an Awadhi word with its Hindi meaning. Figure 4b shows the results after integration of Awadhi word to the Hindi synset along with an AWD tag, showing membership to the Awadhi dialect.

3.3 Knowledge Representation

This section documents the different data representations we developed out of the existing data.

² archive.org/details/in.ernet.dli.2015.481490/mode/1up

³ https://bit.ly/2ySzxWY

⁴ https://pypi.org/project/google-cloud-vision/

⁵ https://regex101.com/

लड़कपिल्ली, "वि० पु० चिबिल्ला लड़का ; वै० लड़खड़ाब किर० स० हिलकर गिरने लगना ;
लड़हरा, सं० पुं० घरी का लंबा पेड़ ।
लड़ाइब, दे० लड़य । ।
लड़ाई, "सं० स्त्री० युद्ध , झगड़ा करब , होय ।"
लड़ाका, वि० झगड़ालू ।

Figure 3: Regex matches as seen on Regex 101 web tool

पार सं० पुं० किनारा;-पाइब, जीतना,-करब,-होब;
-लागब, हो सकना;-लागाइब ।

Figure 4a: Awadhi word पार and its Hindi meaning किनारा in Awadhi-Hindi bilingual dictionary

Hindi word/URL	Old Synset	New Synset
किनारा	किनारा, किनार, कोर, सिरा, छोर, उपांत, अवारी, आर, पालि, झालर	किनारा, किनार, कोर, सिरा, छोर, उपांत, अवारी, आर, पालि, झालर, पार(AWD)

Figure 4b: Hindi word (किनारा) along with its old and new synset

3.3.1 Stop Words and POS Tags

Since there is no pre-existing list of stop words for Awadhi, we make our own. We first create a frequency map of all the words in the corpus and sort them in descending order of frequency of occurrence. Words which have a high frequency, and in addition belong to the class of determiners, prepositions or conjunctions such as यह (this), उसका (their) and और (and) are added to the list of stop words.

We use the Hidden Markov Model (HMM) to POS tag our precompiled Awadhi dataset. The HMM requires a set of observations and a set of states. Words in a sentence are defined as the observations and the POS tags are the hidden states. The HMM uses a transition probabilities matrix and a conditional probabilities matrix. For a given pair of POSTags, say (ADJ and NP) the transition probability TP is defined by the conditional probability:

$$TP = P(\text{ADJ} | \text{NP}) \quad \dots \text{Eq 1}$$

For a given word a and POSTag, the emission probability EP is defined by the conditional probability:

$$E.P. = P(a | \text{NP}) \quad \dots \text{Eq 2}$$

In order to build the two matrices, we use pre-tagged hindi dataset available from the Universal Dependencies, UD_Hindi-HDTB dataset (Riyaz, Martha, 2009). This dataset consists of close to 2000 POSTagged sentences in the Hindi language. Once we train this model, for a new sentence it uses the pre-built matrices to predict the POSTags.

Figure 5 shows the result of POS tagging a Awadhi sentence using the HMM based POS tagger. The English translation of the sentence is - "Being a minister it becomes his duty to listen to both the parties".

3.3.2 Lexicon

We create a lexicon of concept words (nouns) and preferential word pairs with the help of the POSTagged Awadhi dataset. This lexicon serves as a rich source of conceptually cohesive words to build sentences with improved factual correctness.

Input : परधान होय के नाते उनका दुनहू तरफ कि सुनैक जरूरी रहे .
Output : परधान/PROPN होय/PROPN के/ADP नाते/ADP उनका/PRON दुनहू/NOUN तरफ/ADP कि/SCONJ सुनैक/NOUN जरूरी/ADJ रहै/VERB ./PUNCT

Figure 5: Result of POSTagged Awadhi sentence

We pick up nouns from the POS tagged dataset. We plot a graph of NP (noun phrases) identified, based on their word embeddings. Each NP serves as a node and the edge weight is the inverse of the cosine distance between the word embeddings. We generate clusters of the plotted nodes. A dense cluster signifies a set of nouns which are used together frequently and hence represent a conceptually cohesive set. Thus, we pickup the cluster having the highest number of nouns. These NPs serve as the final set of nouns that are included in our lexicon.

We now build the rest of the lexicon. From the dataset, we first allot each ADJ a proximity score based on the number and the closeness of the selected NP around it. We pick a set of top 'n' unique adjectives, based on their scores. These will now serve as the final set of preferential noun-adjective pairs in our lexicon. We perform the same sequence of steps to pick up preferential noun-verbs, noun-pronoun and verb-adverb pairs in the lexicon.

For example, let the noun word be परकृति (nature) for which the corresponding adjectives are अनगिन्त (limitless), सहज (spontaneous), कृतघ्न (not showing gratitude), ममतामयी (mother's kindness) and स्थायी (fixed, not changing).

3.3.3 Digital Dictionary

The inverse dialect mappings created to enrich the present Indowordnet (refer subsection 3.2.1) also serves as a resourceful bilingual dictionary in digital form for a given dialect word. Using our sentence generation model explained in the next subsection, we further enrich this dictionary with example sentences for each word-meaning pair.

3.4 Sentence Generation using Recurrent Neural Networks

We designed a Recurrent Neural Network (RNN) that helps us in generating meaningful sentences using a dialect word as seed. Since Awadhi is a low resourced language, RNN is seen as a good method for sentence generation (Gandhe, 2014). The sentences serve as exemplary sentences for the newly added Awadhi word to the IndoWordnet. The motivation for using RNN rather than pick up an example sentence from the corpus itself is to be able to generate new sentences that highlight the local cultural aspects of the dialect. This aligns with our objective of preserving heritage and we will address this issue as the next step in our project's roadmap.

Alternatively we also used the N-gram model to generate sentences. This model takes in a set of words and generates a score of each possible permutation based on Markov probability rules (Yadav, 2014). The limitation of this model lies in its prerequisite to provide the entire list of words that the sentence would comprise of. Furthermore, it cannot construct sentences from a single seed word as is possible with RNN. It is noteworthy that even though RNN has been used for sentence construction in Bangla (Islam, 2019) and English (Sutskevar, 2011), it is for the first time that it has been used for sentence generation in a Hindi dialect - Awadhi.

We trained the RNN on the set of Awadhi sentences compiled in our corpus from multiple sources as mentioned in section 3.1. The RNN aims at understanding the syntactic constructs of words in a sentence so that it can use this knowledge in predicting words that are most probable in a given context.

To ensure that the sentences being generated are semantically correct, we make use of the preferential word pair lexicon we developed in subsection 3.3.2. During each step of next-word prediction in RNN, the model returns an array of probabilities for the next word. Using the lexicon relations we selectively nullify the probability scores of unrelated words. For example, for the root word - पिता, the top 5 words with highest probabilities in the probabilities array returned by the RNN are [जी, जान, कठोर, अपनी, पिरय]. The noun-adjective lexicon pair for पिता is [पुलकित, परम, कठोर, साध्वी, पिरय]. Hence, after nullifying the probabilities of the words not present in the lexicon, the top 5 words now are [कठोर, पिरय, गुरु, तुल्य, परम].

Our training set consists of “s:t” pairs that correspond to a list of 5 words in sequence and the next-word in sequence respectively. Figure 6 below shows (s) as an array of 5 words in sequence and (t) as the next word in this context.

```
['\n', 'हमरे', 'लिए', 'इतनी', 'जगह'], next_word: काफी  
['बहिका', 'रंग', 'ओ', 'मुखादि', 'फूलमती'], next_word: कि  
['बने', 'सेनी', 'अत्ती', 'तपनि', 'मैहा'], next_word: खुब  
['तेरह', 'मा', 'यहि', 'पोटर्ल', 'पर'], next_word: जेतना
```

Figure 6: Training pairs generated from Awadhi data-set available

To illustrate the process, consider the first training pair. For 5 words - i. \n (newline) ii. हमरे (us) iii. लिए (for) iv. इतनी (this) v. जगह (place) - in sequence the next word is काफी (enough). This “s:t” pair has been extracted from the sequence - \n हमरे लिए इतनी जगह काफी है । (This place is enough for all of us.)

In the fourth training pair, for 5 words - i. तेरह (thirteen) ii. मा (in) iii. यहि (this) iv. पोटर्ल (portal) v. पर (on) - in sequence the next word is जेतना (specific). This “s:t” pair has been extracted from the sequence - कुछ ब्यस्तता के चलते दुइ हजार तेरह मा यहि पोटर्ल पर जेतना काम हुवै क रहा, नाय होइ पावा । (Specific work on this portal couldn't be completed due to some busy schedules in two thousand thirteen.)

Of the available Awadhi sentences for training purposes we use 20% of the dataset for the purposes of validation and the rest for training. Awadhi as a dialect is low resourced and most of the resources available online overlap in their content. During training, we allowed overfitting of the model over the consolidated training set of Awadhi sentences. We observe that overfitting of the data actually helps us to retain the semantic and syntactic relationships between words in the way they occur in the actual text. However, this also leads to a decrease in the overall generalizability of the process of sentence generation.

3.5 Evaluation of Sentence Quality - BLEU

BLEU (BiLingual Evaluation Understudy) is an algorithm for evaluating the quality of machine-translated text from one natural language to another (Kishore, 2002). The BLEU score has been used for measuring machine translated English to Hindi sentences (Malik, 2016). It can also be used for evaluation of sentence similarity.⁶

For each sentence that is generated by the RNN model for a given root word, we create a reference set using the consolidated Awadhi dataset. The reference set selectively contains only those sentences which contain the root word. Choosing sentences which contain the root word ensures that only relevant sentences are compared against and this decreases the chances of getting low scores.

The BLEU model now evaluates the cumulative n-gram scores of the candidate sentence with respect to the reference set, at all orders from 1 to n. NLTK's BLEU model by default calculates the 4-gram cumulative score, with n being set to 4 and the default weights being (0.25, 0.25, 0.25, 0.25). The algorithm finally returns the weighted geometric mean score for all n-gram scores. We make use of this model to evaluate the quality of our generated sentences⁷.

An example is shown in figure 8, where the awadhi root word is जंगल (forest) and the RNN generated sentence is जंगल (forest) मा (in) एक (one) जगह (place or area) आम (mango) पाक (ripe) रहा. (There is one place in the forest where mangoes are ripening). The reference set

⁶ bit.ly/2V640ms

⁷ bit.ly/3aadZLC

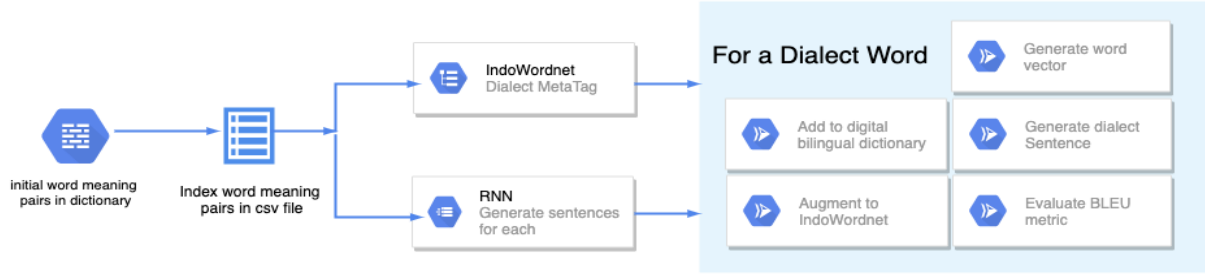


Figure 7: Complete Workflow

Figure 7 shows the complete workflow, starting from word meanings in *Awadhi Shabdkosh*, inverted word-meaning pairs, integration to IndoWordnet to sentence generation using RNN and sentence evaluation using BLEU

consists of 9 sentences which contain the awadhi word - *जंगल*.

The BLEU score is calculated for the above sentence using the reference set giving us a score of 0.4012.

```
example_reference = [
    "घउगङ्गा जंगल की राह लिहेस",
    "जंगल मा बड़ठा रहई",
    "सिकार उठा जंगल मा भाग",
    "सब केउ जंगल मा जुटि के गयेनि",
    "साधू जायि के उही जंगल मा फेरि राम राम करइ लागेनि",
    "जब लगे गयेनि तउ देखेनि कि हारमती जंगल के बन्दी खाने मा परी बिलपत बा",
    "वहि कयि रोउब सुनि के जंगल कयि चिरई जुटि गई",
    "सोना बहिनी रोवत की बीच जंगल मा बयिठी",
    "सोना घरे से जायि के जंगले मा रोवयि लार्णी"]

generated_sentence = "जंगल मा एक जगहाँ आम पाक रहा "

sentence_bleu(example_reference, generated_sentence)

0.40126711450090535
```

Figure 8: BLEU score for the generated sentence

4. Experimental Results

4.1 IndoWordnet Enrichment

The Awadhi-Hindi bilingual dictionary has 37 alphabets ranging from अ to य. Table 3 shows the total number of Awadhi words under column 3 (TW) that exist under each alphabet - shown under column 1 (CE) in English and column 2 (CH) in Hindi. The total number of Awadhi words having one word Hindi meanings are shown under column 4 (TWSM). The inversion process led to an average of 48.48% loss in Awadhi words collected.

The number of Hindi synsets enriched with their Awadhi equivalents and their exemplary sentences due to the next step of integration to the IndoWordnet is shown under column 5 (SE). This step incurs a further miss rate of 30.91% on an average. This loss was seen to occur due to the following two factors - 1) OCR does not identify the Hindi word in the bilingual dictionary correctly. For example, in figure 9 the OCR interprets खट्टापन (sourness) as खापन (no such word exists). The target Hindi word doesn't exist in the IndoWordnet. For example, the Hindi word मइजिल doesn't exist in IndoWordnet. Figure 10 plots the alphabetical inverse mapping losses and IndoWordnet integration losses.

The IndoWordnet consists of 26,000 synsets for the Hindi language⁸ and the number of synsets enriched due to the Awadhi corpus is 3036 (11.68%). This is significant, keeping in mind the scarcely available Awadhi datasets. We believe this number will increase when we proceed with other dialects of Hindi such as Braj, Rajasthani, Marwari etc.

खटासि संस्त्री० खट्टापन, थोडी खटाई, वैस ।		
1	word	meaning ।
48	खटासि	संस्त्री० खापन , थोडी खटाई , वैस ।

Figure 9: (a) Awadhi word खटासि and its meaning in *Awadhi Shabdkosh*. (b) OCR interprets खट्टापन (sourness) as खापन (no such word exists).

4.2 BLEU Scores

The threshold score of 0.6351 was decided on the basis of statistics observed over 158 sentences generated, two for each of 79 Awadhi words chosen randomly from our corpus. The statistics are shown in Table 1. We decided to include sentences with BLEU scores above that corresponding to the 75th percentile score.

Max	Min (Non-zero)	Mean
0.9036	0.1119	0.4679
Median	75th Percentile	90th Percentile
0.4324	0.6351	0.7174

Table 1: BLEU Scores

RNN has not been used yet for generating sentences in Awadhi or any other dialects of Hindi. We show exemplary sentences in Table 2 with scores above threshold for the Hindi words माई (mother) and बच्चा (child) along with their English translation.

The english translations were performed manually by us using the awadhi-hindi and hindi-english bilingual dictionaries. It was seen that for a threshold of 0.6351 sentences were syntactically and semantically correct.

⁸ <https://bit.ly/2XB07HW>

<u>sentences for माई</u>	<u>BLEU</u>
माई (mother) तोहरे (your) संकोच (hesitation) के मारे कुछ (nothing) बोली (spoke) ना (your mother spoke nothing out of hesitation)	0.7788 ✓
माई संस्कार कहां मिली कि हमरे माई केहू खुशी नाहीं देखे नाहीं तौ भर जाय के धारि अपने साथ चली जा	0.2069 ✗
<u>sentences for बच्चा</u>	<u>BLEU</u>
बच्चा (children) लोग जौ खाना (food) लाये (brought) हैं ऊ खाय (eating) लागै (The children started eating the food they had brought with them)	0.8948 ✓
बच्चा (child), मंत्र (mantra) एककइ (once) बार काम(works) करा थइ (Listen child , this mantra works only once)	0.6514 ✓

Table 2 : RNN generated sentences

We made a visualisation tool for generating new sentences and several other tasks mentioned in section 4. The link is added in the references.⁹

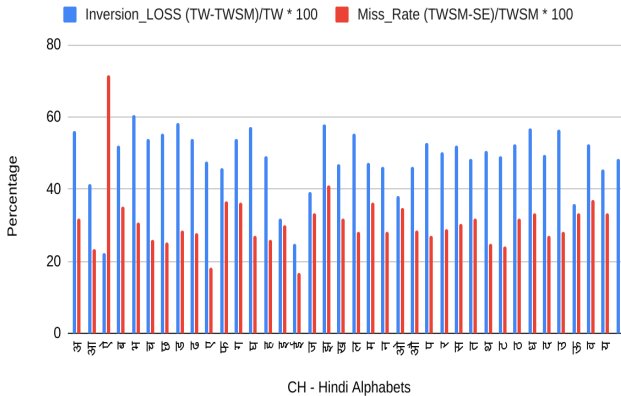


Figure 10: Histogram showing inversion loss and miss percentage.

5. Conclusion and Future Scope

We formulated and presented a methodical and scalable approach to enrich the IndoWordNet. This not only enhances the Indo Wordnet's viability as a social project but also protects local dialects from fading. Similar to the Awadhi dialect, we also plan to expand our approach to the Braj and Marwari dialect. We have bilingual dictionaries for both the dialects. However, lack of a corpus for these dialects is a constraint. Our model needs nothing more than the corpus text and a binary

CE	CH	TW	TWSM	SE
A	अ	552	242	165
Aa	आ	80	47	36
Ai	ऐ	9	7	2
B	ब	817	391	254
Bha	भ	288	114	79
Ca	च	437	201	149
Chha	छ	177	79	59
Da	ड	161	67	48
Dha	ढ	78	36	26
E	ए	21	11	9
Ea	फ	201	109	69
Ga	ग	553	254	162
Gha	घ	163	70	51
Ha	ह	297	151	112
I	इ	44	30	21
Ii	ई	8	6	5
Ja	ज	322	196	131
Jha	झ	105	44	26
Kha	ख	365	194	132
L	ल	272	121	87
Ma	म	578	305	194
Na	न	335	180	129
Q	ओ	42	26	17
Qu	औ	13	7	5
Pa	प	657	310	226
Ra	र	251	125	89
Sa	स	815	391	272
T'a	त	338	175	119
T'ha	थ	65	32	24
Ta	ट	130	66	50
Tha	ठ	86	41	28
Thha	ड	132	57	38
Ttha	द	375	189	138
U	उ	173	75	54
Uu	ऊ	14	9	6
Ya	य	40	19	12
Y'a	य	33	18	12
Total		9027	4395	3036

CE: hindi alphabet in english, CH: hindi alphabet, TW: total Awadhi words in the dictionary for a given alphabet, TWSM: total Awadhi words with single word Hindi meanings, SE: Hindi synsets enriched in IndoWordnet

Table 3: Indowordnet Enrichment statistics

⁹ <https://bit.ly/2K45V19>

mapping of dialect words to the mother language; today's technological armoury is such that if such text is collected in any digital form (textual/ visual/ audio /video) the model can work through it.

Although we achieved encouraging results there are certain shortcomings which if taken care of can make this model more reliable.

Another improvement can be incorporated by designing stemmers for a given dialect. Right now the inverse mappings from bilingual dictionaries contain mappings of 'word-to-word' form but 'phrases-to-words' and 'n-grams' can be considered further. Through 'phrases-to-word' mappings we can decrease the inversion loss percentage of 48.48. Also, sentences generated by our model use the forward probability of words in a sentence. For capturing the complete context, backward probability can give better results. Overall, the model has good potential for further growth. Each dialect of Hindi has its own geographical style and culture. Our future aim would be to generate sentences using RNN that highlight these cultural aspects.

6. Bibliographical References

- Arun Baby, Anju Leela Thomas, Nishanthi N L, and Hema A Murthy, "Resources for Indian languages", CBBLR workshop, International Conference on Text, Speech and Dialogue. Springer, 2016.
- Bhattacharyya P (2010) IndoWordNet de Melo G, Weikum G (2012) Constructing and utilizing WordNets using statistical methods. *Lang Resour Eval* 46(2):287-311
- B.L. Sukhwai, *Modern Political Geography of India*, Stosius Inc/Advent Books Division, ... In the Hindi heartland ... (1985)
- E. G. Caldarola and A. M. Rinaldi, "Improving the Visualization of WordNet Large Lexical Database through Semantic Tag Clouds," 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, 2016, pp. 34-41.
- Gandhe, Ankur, Florian Metze, and Ian Lane. "Neural network language models for low resource languages." *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.
- Islam, Md Sanzidul, et al. "Sequence-to-sequence Bangla sentence generation with LSTM Recurrent Neural Networks." *Procedia Computer Science* 152 (2019): 51-58.
- Khodak, Mikhail, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. "Extending and Improving Wordnet via Unsupervised Word Embeddings." arXiv preprint arXiv:1705.00217 (2017).
- Kishore, Papineni, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
- Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, Fei Xia. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure. In the Proceedings of the 7th International Conference on Natural Language Processing, ICON-2009, Hyderabad, India, Dec 14-17, 2009.
- Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.
- P. Malik and A. S. Baghel, "An improvement in BLEU metric for English-Hindi machine translation evaluation," *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Noida, 2016, pp. 331-336.
- Ritesh Kumar, Bornini Lahiri, Deepak Alok, Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic identification of closely-related Indian languages: Resources and experiments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).
- Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. The Hindi/Urdu Treebank Project. In the Handbook of Linguistic Annotation (edited by Nancy Ide and James Pustejovsky), Springer Press
- Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. "Generating text with recurrent neural networks." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011
- Taghizadeh, Nasrin, and Hesham Faili. "Automatic wordnet development for low-resource languages using cross-lingual WSD." *Journal of Artificial Intelligence Research* 56 (2016): 61-87.
- Varun Gandhi (2018), <https://economictimes.indiatimes.com/blogs/et-commentary/preserving-indias-endangered-languages/>
- Yadav, Arun Kumar and Samir Kumar Borgohain. "Sentence generation from a bag of words using N-gram model." *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies* (2014): 1771-1776.