Evaluating Two Annotated Corpora of Hindi Using a Verb Class Identifier

Neha Dixit
Centre for Technology Studies,
MGAHV, Wardha
bhunehadixit@gmail.com

Narayan Choudhary
ezDI, LLC.
choudharynarayan@gmail.com

Abstract

[In the past few years, Indian languages have seen a welcome arrival of large parts of speech annotated corpora, thanks to the DIT funded projects across the country. A major corpus of 50,000 sentences in each of the 12 of major Indian languages is available for research purposes. This corpus has been annotated for parts of speech using the BIS annotation guideline. However, it remains to be seen how good these corpora are with respect to annotation itself. Given that annotated corpora are also prone to human errors which later affect the accuracies achieved by the statistical NLP tools based on these corpora, there is a need to open evaluation of such a corpus. This paper focuses on finding annotation and other types of errors in two major parts of speech annotated corpora of Hindi and correcting them using a tool developed for the identification of verb classes in Hindi.]

1 Introduction

This paper emerges from a task meant to automatically identify the syntactico-semantic class of Hindi verbs occurring in a sentence. A verb class identifier was developed that took the parts of speech annotated text as input and identified the class of the verbs marked as main verbs in the text. Section 1 and 2 details the development of this automated identifier. This tool was run against two corpora, first the Hindi corpus developed by Microsoft Research India (MSRI) (Bali et al., 2010) and distributed by the Linguistic Data Consortium (LDC)[1] and second, the Hindi corpus developed under the consortia project called Indian Language Corpora Initiative (ILCI) (Choudhary et al., 2011) and distributed by the TDIL[2]. While the MSRI corpus is annotated in the IL-PoST framework (Baskaran et al., 2008) of parts of speech annotation, the ILCI corpus is annotated using a tagset now commonly known as the BIS (Bureau of Indian Standards) tagset. Both of these tagsets are conceptually hierarchical and therefore have top level categories for each of the classes of words. For verbs, both of these tagsets categorize them as either main verb or auxiliary verb. While the IL-POST also includes the morphological information for each of the verbs annotated, the BIS tagset requires only the top level categorization of main or auxiliary verb. We show that both of these corpora have a high number of errors of both omission and commission which should be taken care of before these corpora are used as gold data for further NLP tasks such as statistical parts of speech tagging and so on.

The second section gives an overview of the verb classification used to develop the verb class identifier followed by the third section detailing the development of the knowledge base. The fourth and fifth sections detail the ambiguities arising out of the use of a knowledge base for verb classification and provide a solution for the most frequent cause of the ambiguities. The last sections present the results achieved after evaluating the verb class identifier against the LDC and the ILCI corpora followed by an error analysis.

[1] https://catalog.ldc.upenn.edu/LDC2010T24
[2] http://tdil-dc.in

2 Verb Classification

While we base the classification of verbs on the traditional classification into transitive and intransitives, we extend the category into multiple sub-classifications. Verbs are classified into a total of 13 categories. The classification we use emanates from the practical use envisaged for such a knowledge base. While the major categories are traditionally four in number (namely, intransitive, transitive, causative and double causative), we further classify the intransitives into 7 sub-classes based on some diagnostic tests which govern their syntactic function and affect or validate what constructions they allow in a sentence.

2.1 Classification of Intransitive Verbs

There are three diagnostic tests applied to classify the intransitive verbs. These diagnostic tests are as follows.

2.1.1 Allows ergative –ne

We know that some verbs require the marking of ergative case marker –ne in the simple past sentences in Hindi while with the others the same is not allowed. While this is always required with the transitive, causative and double causative verbs, the same does not always happen with the intransitive verbs.

For example, in the following two sentences where the main verb is intransitive, the use of –ne makes the sentence ungrammatical:

*Hindi	मोहन	ने	बहुत	सोया
IPA:	mohən	ne	bahut	soja:
Gloss:	Mohan	ERG	very	sleep-PST-MSG
Meaning:	Mohan slept a lot			

*Hindi	सुमन	के	सिर	ने	बहुत	चकराया
IPA:	sumən	ke	sir	ne	bahut	cəkra:ja:
Gloss:	Suman	POS	head	ERG	a lot	reel
Meaning:	Suman's head reeled a lot.					

But in the sentences below where again the main verb is intransitive, the use of –ne marking is perfectly fine:

Hindi	शेर	ने	जोर	से	दहाड़ा
IPA:	sher	ne	ʤor	se	dəha:ɖa:
Gloss:	lion	ERG	strong	ASSOC	roar-PST-MSG
Meaning:	The lion roared strongly.				

Hindi	राम	ने	जोर	से	छींका
IPA:	ra:m	ne	ʤor	se	$c^hĩ$:ka:
Gloss:	Ram	ERG	strong	ASSOC	sneeze-PST-MSG
Meaning:	Ram sneezed heavily.				

2.1.2 Allows Adjectival Use of the Perfective Form

All the intransitive verbs in their perfective forms cannot be used as an adjective. While some verbs allow this, others do not. For example, in the following sentences, the perfective forms of the verbs are being used as adjectives to modify nouns or noun phrases following it:

*Hindi	तीर्थ	यात्रा	पर	गए	हुए	लोगों	की…
IPA:	ti:rt̪^{h}	ja:tra:	pər	gəe	hue	logon	ki:…
Gloss:	pilgrimage		on	go-PFT	be-PFT	people	of
Meaning:	People who have gone on a pilgrimage…						

Hindi	संत	गिरे	हुए	लोगों	को	उठाते	हैं
IPA:	ənt̪	gire	hue	logon	ko	ut̪^{h}ate̪	hɛ̃
Gloss:	saint	fall-PFT	be-PFT	peo-ple	ACC	pick-IMPF	be-PL
Meaning:	Saints help the fallen people.						

Hindi	झाड़ी	में	अटका	गेंद	खो	गया
IPA:	ʤ^{h}aɖi:	mẽ	ət̪ka:	gẽd̪	k^{h}o	gəja:
Gloss:	bush	LOC	stick-PST	ball	lose-PST	go-PST
Meaning:	The ball stuck in the bush got lost.					

But if we use the same form of some other verbs as adjectives, the sentence becomes ungrammatical or sounds awkard:

*Hindi	कांपा	हुआ	लड़का	गिर	गया
IPA:	kã:pa:	hua:	ləɖka:	gir	gəja:
Gloss:	shiver-PST	be-PST	boy	fall	go-PST
Meaning:	The boy, who had shivered, fell.				

*Hindi	घाटी	में	चीखा	लड़का	मर	गया
IPA:	g^{h}a:ʈ	mẽ	ci:k^{h}a:	ləɖk	mə	gəja:

	i:			a:	r	
Gloss:	val-ley	LOC	shriek-PST	boy	die	go-PST
Mean-ing:	The boy, who had shrieked in the valley, died.					

*Hindi	दौड़ा	हुआ	लड़का	जीता
IPA:	dəuɖa:	hua:	ləɖka:	ʤi:ʈa:
Gloss:	run-PFT	be-PFT	boy	win-PFT
Meaning:	The boy, who had run, won.			

2.1.3 Passivization Selection

Similar to other constraints, some intransitive verbs does not allow passivization while other intransitives do. For example, in the following sentences the passive use of the main verb is alright:

Hindi	अब	उठा	जाए
IPA:	əb	uʈʰa	ʤae
Gloss:	now	stand-PFT	go-OPT
Meaning:	Let's stand up now.		

Hindi	शेर	से	गुर्राया	नहीं	गया
IPA:	ʃer	se	gurraja:	nəhĩ:	gəja:
Gloss:	lion	ASSOC	roar-PFT	NEG	go-PFT
Meaning:	The lion could not roar.				

But the same does not hold true for the intransitive verbs as used in the following sentences:

Hindi	अब	उजड़ा	जाए
IPA:	əb	uʤɖa:	ʤa:e
Gloss:	now	wreck-PFT	go-OPT
Meaning:	Let's get wrecked now.		

Hindi	उससे	घबड़ाया	नहीं	गया
IPA:	usəse	gʰəbɖa:ja:-PFT	nəhĩ:	gəja:
Gloss:	he-DAT	bewilder-PFT	NEG	go-PFT
Meaning:	He could not get bewildered.			

Taking these three diagnostics test as the base of the sub-classification within intransitive verbs, we come with a total of 13 classes of verbs as noted in the table below:

Verb Class	Verb Class Label
Causative	CAUS
Copular Verb	COP
Second Causative	DB_CAUS
Intransitive (+Adjectival)	INTR_ADJ
Intransitive (+Adjectival +Passivization)	INTR_ADJ_PAS
Intransitive (+Ergative)	INTR_ERG
Intransitive (+Ergative + Adjectival)	INTR_ERG_ADJ
Intransitive (+Ergative + Adjectival +Passivization)	INTR_ERG_ADJ_PAS
Intransitive (+Ergative +Passivization)	INTR_ERG_PAS
Intransitive (+Passivization)	INTR_PAS
Intransitive	INTR
Intransitive/Transitive	TRAN_INTR
Transitive	TRAN

Table 1: Classification of Verbs

3 Developing a Verb Class Knowledge Base

Identification of verbs and their classes bases itself mainly on what we consider the largest ever knowledge base of Hindi verbs collected from various sources such as dictionaries, corpus and others. The knowledge base contains a total of 3240 verbs and all of their morphological forms, including spelling variations and common mistakes. The morphological forms and the spelling variations are further given their own labels in the knowledge base itself along with the class for each of the verbs and their forms. The structure of knowledge base has been illustrated in the following table:

ID	Word	morph_type	verb_class
86727	पीना	inf_msg	Transitive
86728	पीनी	inf_fsg	Transitive
86729	पीने	inf_pl	Transitive
86730	पीता	impf_msg	Transitive
86731	पीती	impf_fsg	Transitive
86732	पीतीं	impf_fsgh	Transitive
86733	पीते	impf_pl	Transitive
86734	पिया	pft_msg	Transitive
86735	पियी	pft_fsg_var1	Transitive

Table 2: Structure of Knowledge Base

With inclusion of 3240 verbs, we get a total of 149,518 words present in the knowledge base.

4 Types of Ambiguities in Verb Classification

The knowledge base forms as main base for the assignment of verb classes in a given sentence. However, the knowledge base itself cannot cover all the cases as there are verbs which can fall into more than one class depending on the context in

which they appear. An analysis of the verbs present in the knowledge shows that it has 1260 words that occur twice in the corpus. These words are possible causes of ambiguity resulting into incorrect assignment of class to the verbs. Our analysis shows that these ambiguities could be of 5 different types as mentioned below.

4.1 Ambiguity: Verb Root vs. Perfective Verb

This occurs mainly because of a derivational process used in Hindi and several other Indian languages where a valency is added by vowel lengthening. For example while the verb जग/ʤəg, an intransitive verb, means "to wake up" the verb root जगा/ʤəga, a transitive verb, means "to awaken". While जगा/ʤəga is a verb root, it is also the perfective inflectional form of the verb root जग/ʤəg and this way जगा/ʤəga gets two verb classes which need to be resolved.

While a solution to disambiguate this type of ambiguity has been implemented as described in by analysing the verb group patterns (as mentioned in the section below), the other types of ambiguities (described below) have to be taken care of at the word level.

4.2 Ambiguity: Conjunctive Participle vs. Perfective Verb

Another type of ambiguity which has a chance of becoming frequent if the genre of the corpus under test is of non-formal kind is that conjunctive participles can get confused with the perfective of the verbs ending on consonant -क/-k. Conjunctive participles are usually formed by adding the auxiliary verb –कर/-kər to any verb root and give a sense of perfective aspect to the verb. While the formal way of creating the conjunctive participle is to either attach -कर-/kər to the verb root itself or juxtaposing it afterwards, informally the variant -के/-ke is used. Thus we can have खाकर/kʰa:kər and खाके/kʰa:ke having the same sense and used interchangeably. Except for a few frequent use of this variant such as खाके//kʰa:ke, रोके/roke, कसके/kəske, etc. most of the time this variant is not used. And this is why we have ignored finding out a rule-based solution to disambiguate this.

4.3 Ambiguity: Perfective Verb vs. Infinitive Verb

Some verbs that end with consonant –न/-n as in छान/cʰa:n, मान/ma:n, जान/ʤa:n etc. may be sharing the same grapheme and may be homophonous with some other verb's infinitive form. Thus a verb like छाना/cʰa:na: may have two meanings, the first being the perfective of the verb root छान/cʰa:n (to filter) and another as the infinitive form of the verb root छा/cʰa: (which means "to cover the roof"). However, this type of ambiguity is also limited and count only 3 in Hindi. For this very reason, we have also ignored disambiguating this for the time being.

4.4 Ambiguity: Perfective Verb vs. Imperfective Verb

There are also a couple of verbs which can be interpreted as imperfective of a verb root and perfective of another. There are two verb root pairs that create this problem. The first pair is जीतना/ʤi:tna: and जीना/ʤina:, meaning respectively "to win" and "to live". The second pair is बरतना/bərətna बरना/bərna, meaning respectively "to follow" and "to choose".

4.5 Ambiguity: Verbs in multiple classes

While it is very common in other languages such as English that the same verb is used both as transitive and intransitive, the same is very rare in Indo-Aryan languages like Hindi. Out of all the verbs that we have analysed, we found only one verb that can be used both as transitive and intransitive. This verb is ऐंठना/ɛ̃ʈʰna: (meaning "to writhe" or "to snatch by deceit").

5 Root vs. Perfective: Disambiguation

Taking a cue from the work done on identification of verb groups in Hindi by Choudhary et al. (2011a), we perform an analysis of the total verb group templates as defined here. Choudhary identifies a total of 675 templates covering all the verb groups possible in Hindi, including the compound verb constructions. As we know that for each of the verb groups found in Hindi, the class is defined by the main verb and this main verb occurs at the start of the verb groups. If we know the morphological type of the main verb (knowing whether it is verb root or a perfective form), we can identify the class of the verb with the help of the verb groups.

An analysis of the 675 verb group templates shows that there are only 30 such templates where both verb root (VR) and perfective verb (VR_pft) can stay. This means we basically have to find out the disambiguation rules for only these verb groups. The rest are taken care of automatically as proper, grammatical structure of Hindi would not allow them. These verb group templates are noted in the table below:

Template with First word as either VR or VR_pft	Ambiguity
VR_pft+ja+rah_pft+ho_fut	Yes
VR_pft+ja_pft+ho_impf+prs_aux	Yes
VR_pft+ja+rah_pft+prs_aux	Yes
VR_pft+ja_inf+cahiye+pst_aux	Yes
VR_pft+ja+rah_pft+pst_aux	Yes
VR_pft+ja_inf+cahiye	Yes
VR_pft+ja_impf+prs_aux	Yes
VR_pft+ja_pft+prs_aux	Yes
VR_pft+par_pft+prs_aux	Yes
VR_pft+ja_pft+pst_aux	Yes
VR_pft+par_pft+pst_aux	Yes
VR_pft+ja+rah_pft	Yes
VR_pft+ja_fut	Yes
VR_pft+ja_impf	Yes
VR_pft+ja_opt	Yes
VR_pft+ja_pft	Yes
VR_pft+rah_pft	Yes
VR_pft+VINF	Yes
VR_pft+VINF_imp	Yes
VR_pft+ho_fut	No
VR_pft+ho_impf	No
VR_pft+ho_opt	No
VR_pft+ho_pft	No
VR_pft+kar_fut	No
VR_pft+kar_imp	No
VR_pft+kar_impf	No
VR_pft+rah_fut	No
VR_pft+rah_imp	No
VR_pft+rah_opt	No
VR_pft+rakh_pft	No

Table 3: Possible Ambiguous Verb Group Templates

Now, if we closely analyse these auxiliary verbs in the given templates, we find that these auxiliaries actually do not allow perfective verbs to occur as their main verbs. This conclusion is based on a corpus study done on the EMILLE (McEnery et al., 2000) corpus and the Gyan-Nidhi corpus as well as some exact searches done on a prominent search engine. For example, in the phrase जगा जाएगी/ʤʒəga: ʤa:egi:, the main verb जगा/ʤʒəga: or for that matter any other verb can never be inflected for the perfective aspect. Similar is the case with the place of other main verbs in the templates having ambiguity (noted with "yes" in table V above).

6 Evaluating against Corpora

The tool was given two corpora as input– the LDC and the ILCI. A summary of the error analysis on the results achieved has been provided here.

6.1 The LDC Corpus

LDC corpus contains a total of 4839 sentences annotated in the IL-PoST framework. When run against the verb class identifier, we get the following results:

Total Main Verbs Found	8,386
Total Main Verbs Classified	8,048
Unclassified Main Verbs	338
Error Percentage	4.030527

The 4% of error in identifying a verb class for a verb marked as VM emerges due to four different reasons as noted below:

Error Types	Frequency
Spelling Error	258
Annotation Error	39
Tokenization Error	18
Echo-Words	23
Total Errors	338

6.2 The ILCI Corpus

The Hindi corpus of ILCI contains 50,000 sentences from two domains of health and tourism. Summary of the error analysis done on the output of this corpus is given below:

Total Main Verbs in ILCI	87,801
No. of Main Verbs Classified in ILCI	82,232
No. of Main Verbs Unclassified in ILCI	5,569
Accuracy on the ILCI Corpus	93.66
Error Percentage	6.34

Further analysis breaks down the types of errors found in this evaluation. This has been shown in the table below:

Error Type	Tourism	Health	Overall	% of Errors
Annotation	3396	1721	5117	0.91604

Error				
Echo word	104	144	248	0.044397
Spelling Error	31	96	127	0.022735
KB Error	74	8	82	0.01468
Pre-Processing Error	3	9	12	0.002148

Table 4: Error Types in the ILCI Corpus

7 Error Types

Evaluating against the ILCI corpus and following the error analysis, we find that there are four types of errors in the annotated text. "KB errors" are errors of knowledge base used in the verb classification tool itself. A summary of the errors for each of the tokens at the annotation level are shown below:

Correct Tag for VM	Unique Words	Frequency of Errors
JJ	420	3,543
NN	660	1,868
Echo	120	254
KB	41	213
Spelling	109	127
Tokenization	26	30
PP	7	28
RB	6	11
RPD	2	9
DMD	5	5
FW	1	1
PRP	1	1
CCD	1	1
DMR	1	1

Table 5: Error Types and their Frequency in the ILCI Corpus

The four major types of errors are mentioned below.

7.1 Annotation Error

Annotation errors are the errors in the assignment of the parts of speech tags to the text. As seen in the table above, the highest number of words marked incorrectly as main verb adjectives which should have been marked as JJ. This is followed by words that should have been marked as NN but are marked as VM. Some prepositions, adverbs, particles and demonstratives are also marked as VM. Some examples such errors are noted below:

Actual Tag	Example Words
JJ	स्थित, पैदा, उपलब्ध, पता, स्थापित, प्राप्त, खड़े, प्रदान, घिरा, शुरू, तैयार, सेवन, निर्मित
NN	सेवन, प्रयोग, आराम, नजर, ध्यान, लाभ, काम, निर्माण, मालिश, याद, इस्तेमाल, बढ़ावा, बढ़ावा
PP	पर, बाहर, सामने, पास, अंदर, आर-पार, का, पारकर
RB	खासकर, नहीं, रूबरू, ऐसे, खड़े-खड़े, सिर्फ
RPD	वाला, ही
DMD	इससे, यह, यहाँ, वहाँ, वहीं

Table 6: Examples of Annotation Errors

7.2 Echo word

Echo-words are a type of reduplication used as a method of word formation for emphasis and other semantic purposes. Using the echo-formation as the word formation process, a non-word is used together with the actual word to add a meaning to it. This phenomenon is seen all the content class words of Indian languages.

However, when it comes to be captured at the level of language computation, this has not been covered yet in most of the cases. The current knowledge base used in the verb class identifier also does not cover the echo-words. Therefore, the main verbs when used in their echo-formation forms do not get detected. Some examples of such words as shown in the examples below:

खाने-पीने, चलने-फिरने, कूट-पीसकर, आने-जाने, घूमने-फिरने, आना-जाना, चलते-चलते, पहुँचते-पहुँचते, उठने-बैठने, मिलता-जुलता

7.3 Spelling Error

The ILCI corpus contains text that is usually corrected for any spelling errors. But some errors have still been found in our analysis. Some of these spelling errors are as follows:

होगें, धों, बढने, भिगों, रखनें, करेगें, करों, चढ़कर, छुपे

7.4 Pre-Processing Error

Parts of speech annotation is usually done after the text is pre-processed and tokenized properly. The same is true with the ILCI corpus as well. However, some errors of pre-processing/tokenization are still left in the text itself. Some examples are shown below:

"(बोलना, ", "बहना'", "``देखने``", "``देखो", "``सुनने``", "``सूँघने``", "आना'", "आने'", "करें-", "काँट", "किया।", "खायें/खिलायें", "जाँचे,"

8 Conclusion

NLP community in India must be elated to have received a big annotated corpus in many Indian languages, including Hindi. These corpora really help a lot in developing next generation of NLP tools for various purposes. However, these corpora are labor intensive tasks and prone to human errors. Errors have been noted in almost all of the human annotation tasks including the Penn Treebank (Manning, C., 2011), the same is true also for other corpora. We have shown here a method to check the accuracy of the tags assigned to main verbs, done the error analysis and pointed out the errors that should be taken care of in the next release of the ILCI corpus and the LDC corpus so that users of the corpora do not need to do the same task again. The verb class identifier tool we used to mark the possible errors can also be used to check the accuracy of any other Hindi corpus annotated for parts of speech tags, thereby alleviating the time taken for error analysis.

Reference

Bali, Kalika, Monojit Choudhury, Priyanka Biswas, Girish Nath Jha, Narayan Kumar Choudhary and Maansi Sharma. 2010. Indian Language Part-of-Speech Tagset: Hindi. Linguistic Data Consortium, Philadelphia.

Baskaran, Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha and K.V. Subbarao. 2008. A Common Parts-of-Speech Tagset Framework for Indian Languages. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Daniel Tapias (Eds.) Proceedings of the Sixth Inter-national Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Choudhary, Narayan and Girish Nath Jha. 2011. Creating Multilingual Parallel Corpora in Indian Languages. In: Proceedings of 5th Language Technology Conference, Poznan.

Choudhary, Narayan, Girish Nath Jha, Pramod Pandey. 2011a. A Rule based Method for the Identification of TAM features in a PoS Tagged Corpus. In: Proceedings of 5th Language Technology Conference, Fundacja Uniwersytety im. A. Mickiewicza, Poznan.

Manning, Christopher D. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In: Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing, 12th International Con-ference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608, Springer