A Rule based Method for the Identification of TAM features in a PoS Tagged Corpus

Narayan Choudhary JNU, New Delhi choudharynarayan@gmail.com

Pramod Pandey JNU, New Delhi pkspandey@yahoo.com

Girish Nath Jha JNU, New Delhi girishjha@gmail.com

Abstract

For a task of natural language understanding, the identification of tense, aspect and mood (TAM) features in a given text is of importance in itself. A closer look at the verb groups in a sentence can give the exact combination of the TAM features the verb group carries. While the verb group consisting of one word could be easily interpreted for the TAM features, the TAM features of the verb groups consisting of more than one word (as witnessed in many languages) can be identified exactly through a rule based method. In this paper we present a rule based method to capture the TAM features denoted by verb groups in Hindi.

Keywords: Verb Group Identification, TAM Identification, tense, aspect, mood, identification, Hindi, Indo-Aryan, Dravidian, Languages, chunking, local word grouping

1. Introduction

The exact information about the tense, aspect and mood (TAM) features can be helpful in many tasks of natural language processing (NLP) including machine translation. The statistical methods of machine translations always rely on a huge data to derive the TAM features encoded in the source language. we present here a rule based method to deduce the TAM features plus a few other information about the structure of the verb groups. The method presented here first looks for the specific patterns of the verb groups in the source language (Hindi is tested here) and then we draw templates for the participants of the verb groups and give them a label (tag) that encodes the TAM features plus any other verb group specific features.

2. Prerequisite Resources

The only prerequisite resource we need is a parts-of-speech (PoS) annotated text as the input. This PoS annotated text should have information about the category of the verb and any other TAM related inflectional marks that the verbal word may contain. We follow the Indian Languages – Parts of Speech Tagset (IL-PoST) annotation framework as described in Baskaran, S. et.al. (2008). We use as test data the Linguistic Data Consortium (LDC) corpus containing above 4800 sentences (above 98 thousand words) prepared in this framework (Bali, K. et.al., 2010). We have also made some modifications in this annotation framework to accommodate more of the TAM markers present morphologically.

3. Verb Groups

By verb groups (VGs) here we mean the word constituents that have a verbal intent in the sentence. In the PoS annotation it must have a verbal mark, either main verb (VM) or auxiliary verb (VAUX). It consists of the serial verb constructions (Aikhenvald, A. and Dixon, RMW, 2006) also as witnessed in many languages. A verb group consists of at least one main verb and can optionally include n-number of auxiliary verbs. There cannot be a verb group consisting only of auxiliary verb(s).

4. TAM Identification

Tense, aspect and mood can be identified with a study of the verb groups of the language concerned. Generally, a descriptive grammar of the language contains all the information about the tenses, aspects and moods that the language can express through the verb groups. While the verb morphology may contain in itself the TAM features, much of the TAM features are expressed periphrastically, through the use of what is commonly called the auxiliary verbs. These auxiliary verbs in any language are limited to some exhaustively countable number of words and they denote some specific tense, aspect or mood features, separately or combined, that cannot be expressed only through morphology.

4.1. TAM Identified through Verb Morphology

The verb morphology of any language may contain inflections for many features besides the TAM and finiteness features that are specific to verbs. For example, a verb may also inflect in agreement with the gender, number and person (GNP) and other features such as honorificity.

In Hindi, a verb root can inflect to generate 25 morphological forms. Not all of these forms encode unique TAM features. However, these 25 forms can be grouped into a number of unique TAM and finiteness features they denote. These TAM and the finiteness features are the ones used in identifying the actual TAM features of the verb groups. Thus the following table shows the different TAM features that can be expressed through the verb morphology of Hindi.

TAM and Finiteness Features	No. of Forms	Co-Occurring Form
imperative/verb root	1	
infinitive	3	
imperfect	4	
perfect	4	optative
future	9	

optative (1sg)	1	
imperative/optative 2sg pl. honorific	1	
imperative	2	

Table 1: TAM and Finiteness Features Expressed through Verb Morphology

As seen in the table above, some TAM and finiteness features are expressed by more than one form. This is because they are inflectionally marked for GNP and have a different form but denote the same TAM feature. There is also one form that is ambiguous: the perfect feature having 4 GNP forms. One of these four perfect forms (plural honorific) can also be used in the optative mood.

We identify these features through the PoS annotation. The IL-PoST framework provides for a hierarchical structure of the PoS features. This ensures that we have all of these morphological features marked at the PoS level.

4.2. TAM Identified Periphrastically

As we can see in Table 1 above, not all the tenses, aspects and moods are covered through the verb morphology. In this section we present the tenses, aspects and moods marked periphrastically.

4.2.1 Tenses Marked Periphrastically

In Hindi only future tense is marked morphologically. Present and past tenses are marked periphrastically (except for the perfect form which is commonly understood as the past marker in Hindi). The copular verb /ho/ 'be' is the only verb which has all the three tense forms- present, past and future. All the other verbs lack the past and present forms at the morphological level. So, when the present and past tenses are to be expressed for the verbs other than 'be', the past and present forms of this copular verb is used. This is just one type of periphrastically formed verb group we find. There are several other TAM features that are marked periphrastically.

4.2.2 Aspects Marked Periphrastically

For the aspects, Hindi has only two aspects that are marked morphologically- perfect and imperfect. For all the other aspects auxiliary verbs are used. These auxiliary verbs are actually the verbs with a lexical meaning of their own. But when they are used as auxiliary verbs, they lose their lexical meaning and give the sense of a different aspect.

The most important among the periphrastically marked aspects is the progressive aspect marker. This aspect is marked with a lexical verb, meaning "live" or "remain" or "stay". The verb form in Hindi is perfective form of the verb /rəh/. When it is used as a progressive aspect marker it loses its lexical meaning.

There are also other aspects that are marked periphrastically. Kachru, Y. (2006) notes four other aspects, namely inceptive, continuative, durative and frequentative. Kachru (ibid) notes that "these are not as general in distribution as imperfect, perfect and progressive.' That is, they have some restrictions and cannot occur as freely as the progressive aspect. There are also some other aspects that share a boundary with the modal category. These are what Kachru (ibid) describes as the presumptive, contingent and past contingent.

However, we find some more of the aspects that can be witnessed in the real world data. Two aspects can get bounded together to denote a new aspect. For example, in the cases when an aspect marking lexical verb is also inflected for the morphological aspect, it can denote two aspects together. For example in the sentence below:

vəh kɪtab pəɽʰne ləg-ta tʰa
he book read.INF apply-IMPF.MSG be.PST.MSG
He used to start reading the book.

The auxiliary verb /ləg/ is the inceptive marker verb while it is in the imperfective form. And thus two aspects (inceptive and imperfective) are marked simultaneously.

The table 2 below shows a summary of all the different aspects that can be identified in Hindi.

Sl. No.	Aspect Name	Label/Tag
1	Perfect	pft
2	Null	0
3	Continuative	cnt
4	Durative	dur
5	Imperfect	impf
6	Imperfect Continuative	impf_cnt
7	Imperfect Durative	impf_dur
8	Imperfect Frequentative	impf_frq
9	Imperfect Inceptive	impf_ince
10	Imperfect Perfective	impf_pft
11	Imperfect-Perfect-Continuative	impf_pft_cnt
12	Perfect Continuative	pft_cnt
13	Perfect Durative	pft_dur
14	Perfect Frequentative	pft_frq
15	Perfect-Imperfect-Continuative	pft_impf_cnt
16	Perfect Inceptive	pft_ince
17	Progressive	prog
18	Progressive Durative	prog_dur
19	Simple	sim
20	Simple Inceptive	sim_ince

Table 2: Summary of Aspects Marked in Hindi

4.2.3 Moods Marked Periphrastically

Like the aspects, there are only two moods that are marked morphologically- imperative and optative. Other moods are marked periphrastically through the help of auxiliary verbs. The declarative is the default mood marker. All the moods that can be marked in Hindi verb groups are summarized in the table below.

Mood Names	Label/Tag
Imperative	imp
Optative	opt
Abilitative	abil
Declarative	dcl
Permissive Imperative	perm_imp
Permissive Optative	perm_opt
Counterfactual	cfct
Desiderative/Suggestive	sugg
Permissive Desiderative	perm_sugg
Optative Probabilitative	opt_prob

Table 3: Summary of Moods Marked in Hindi

5 Drawing Verb Group Templates

Based on the morphological features and the VM or VAUX functionality of verbs, a general template can be drawn to capture the verb groups and then assign a label to them based on the grammatical features they carry. A

verb group contains at least TAM features. It may additionally contain other features such as passive voice construction and compound verb structure. This is true for all the Indo-Aryan and Dravidian languages and may be true for many other languages.

A study of different types of verb group structures in Hindi can be brought down to an abstract level of templates. We did such a study, basing ourselves on the verbal paradigms as shown in various linguistic literatures on the language (Guru, K. (1978); Kachru, Y. (2006)˜ among others).

5.1 Structure of Verb Groups in Hindi

In Hindi (and in all the major Indian languages), a verb group may constitute of one or more verbs. As per the definition provided above in §3, a verb group in Hindi may constitute of just one word and can have at most five words in it. The example sentences i-v illustrate this.

i. One-word VG:-

vəh gəya
he go.PFT.MSG
He went.

ii. Two-word VG:-

vəh ja-ta hɛ
he go.IMPF.MSG be.PRS.SG
He goes.

iii. Three-Word VG:-

vəh gʰər ja rəha hɛ
he home go.VR live-PFT.MSG be.PRS.SG
He is going home.

iv. Four-Word VG:-

vəh kɪtab pəɽʰ-ta ja rəha tʰa
he book read-IMPF.MSG go.VR live-PFT.MSG be.PST.MSG
He keeps on reading the book.

v. Five-Word VG:-

vəh	kɪtab	pəɽʰ-ta	cəl-a	ja	rəha	tʰa
he	book	read-IMPF.MSG	walk-PFT.MSG	go.VR	live-PFT.MSG	be.PST.MSG

He had gone on reading a book (for a long time).

Our study also shows that a verb group can be broken only by what is called the particle. Particles in Hindi constitute of a few indeclinable words. These words in Hindi are the negative markers (/nəhĩ/, /na/, and /nə/) topicalizers (/hi/ "only", /to/ "then/so", and /bʰi/ "also") question words (e.g. /kya/ "what", /kəhã/ "where", /kɔn/ "who", etc.) and a few other words namely /sa/, /si/ /se/ (all meaning "like" in the sense of its conjunctive or quotative function) and /wala/, /wali/, /wale/. The particles /wala/, /wali/ and /wale/ are commonly denoted as the *wala* particle in Hindi. The *wala* particle is multifunctional in Hindi. But when it comes attached with a verb, it has only two functions. It either functions as an agentivizer (example vi) or marks the approximative aspect (example vii).

vi. Agentivizer use of the *wala* particle

kʰane wale ləɽke ko bʊlao
eat-INF AGENT boy.OBL ACC.MSG call.IMP
Call the eating boy!

vii. Approximative Aspect Marking by the *wala* particle

ləɽka kʰane wala hɛ
boy eat-INF APPROX be.PRS.MSG
The boy is about to eat.

When the *wala* particle functions as an agentivizer, it may not be considered as part of the verb group because the function it plays together with the verbal word is that of a noun and there is no verbal intent in it. But when functioning as an approximative aspect marker, it must be included as a part of the verb group because it plays a part in denoting the TAM of the verb group. This information is used in defining and identifying the verb group templates and giving a valence to particles inside them.

With the above description we know that a verb group must contain only verbs and if anything else comes in between it must be the particles as described above. While defining the templates, we ignore all the particles except for the *wala* particle functioning as an aspect marker.

5.2 General Verb Group Templates

With the details above, we defined the patterns of Hindi verb groups. We used the morphological information as represented through the PoS annotation of each of the verbs playing a part in the verb groups along with the VM/VAUX dichotomy as specified in the PoS annotation. These templates would have as many variables as the number of verbs playing a role in verb group. Thus, we can have VG templates that would have just one variable to VG templates having up to five variables. A few examples of the VG templates along with their TAM tags are given in Table 4 below.

VG Template	TAM Tag
prs_aux	VG.prs.sim.dcl
VR_impf+prs_aux	VG.prs.impf.dcl
VR_impf+ban_pft+prs_aux	VG.prs.impf.abil
VR_impf+ja+rah_pft+prs_aux	VG.prs.prog_dur.dcl
VR_impf+cal_pft+ja+rah_pft+prs_aux	VG.prs.prog_dur.dcl
VR_pft+ja_pft	VG.pas_pst.sim.dcl
VR_pft+ja_impf+prs_aux	VG.pas_prs.impf.dcl
VR_pft+ja_inf+lag_impf+prs_aux	VG.pas_prs.impf_ince.dcl
VR+dal_fut	VG.cv_fut.sim.dcl
VR+ja_pft	VG.cv_pst.sim.dcl
VR+dal_pft+ja_opt	VG.cv_pas_0.0.opt
VR+de_pft+prs_aux	VG.cv_prs.pft.dcl

Table 4: Sample VG Templates and TAM Tags

We use variables to define the template. These variables are of two kinds- tag variables and the word variables. Tag variables are the variables the values of which are identified through the PoS tags they have in the input. A word variable is identified by the actual word. For example, the variable VR_impf is a tag variable and is identified by the presence of the tag 'impf' in the annotation of the verb. A word variable is identified by the actual value set for that variable. While we have around a dozen of tag variables, we have word variables running in hundreds. A word variable may have more than one value as its variable. The number of word variables is higher because we use a word variable for all the auxiliaries and their various TAM and finiteness marked forms,.

Using these variables, we define a total of 177 VG templates that cover all of the verb groups in Hindi except the passive constructions and the VGs having a compound verb.

5.3 Passive Constructions

Passive constructions are rather regular in Hindi. It can be viewed as a transformation of what occurs in the active voice (covered under the general VG templates in preceding section). However, not all the general VG templates have a corresponding passive construction. This is because not all the aspects and moods have a corresponding passive construction. For example the sentences in imperative mood cannot transform into passive voice (it changes into optative mood). Our study shows that out of the total of 177 general VG templates, 67 can transform into passive construction resulting into 67 new VG templates. These templates are marked for their passive marking with the prefix of 'pas' in the TAM tag assigned to them.

5.4 Compound Verb Constructions

The compound verb is a pan-Indian linguistic phenomenon (Abbi, 1992) wherein two verbs occur one after the other. In this construction, while the first verb (V1) gives the main meaning to the VG and remains in the root form, the second verb (V2) gets all the inflectional markings and adds a 'shade' into the meaning of V1. These V2s are limited in number though there is no consensus on how many of such V2s should be included in this category. Hook (1974) provides a summary of the various lists of V2s of Hindi suggested by several scholars. In our study, we prepared a list of all such words that can possibly appear even once as a V2 inside a VG and made templates for each of them separately, under the compound verb (CV) category. A total of 61 such V2s are included in this list. This is larger than any other such list in Hindi found in the literature. The VGs containing a CV can be identified with the prefix of 'cv' in the TAM tag assigned to them.

Given that a compound verb may also behave more like a main verb where the V2 takes all the TAM markings, it can be proposed that we simply apply the same templates as drawn for general and passive VG constructions to fit into these compound verbs as well. The only change that will take place in the templates will be the addition of the V1 in its stem form at the beginning. For example for the one word VG templates we will have a two word VG templates which will be a compound verb construction.

While the above method of CV template derivation is helpful in deriving the CV templates, it is not true for all the 244 non-CV VG templates (the general and passive VG templates). A subjective study of the possibility of even one CV construction in these templates was done by us. And we found that there are 40 templates out of 244 non-CV VG templates that will not have a corresponding a CV construction.

There are also some semantic restrictions on the combination of V1 and V2 and not all the V2s can form a compound verb with all V1s (Paul, S., 2006). Nespital, H. (1997) provides a database of Hindi complex predicates including the CV constructions. If one wants to do semantic analysis of the VGs, such a database would be of great help. But we are here concerned merely about the TAM features of the VGs and this we can achieve without such a database.

Based on the description above, we have a total of 204 CV templates which would cover all of the VGs containing a CV. For example for a non-CV template of VR_impf+prs_aux, we can have a corresponding CV template of V1+V2_impf+prs_aux. However, this is only an abstract drawing. If we leave these templates as they are, they will violate the restrictions on the combination of V1 and V2 and will apply to both of them universally. Although there are some templates in which all the V2s can fit in. We have identified a total of 37 such non-CV templates and call them Universal CV Templates. All the other templates (i.e. 204-37=167) templates would have only a few V2s forming a CV template.

Thus we have a total of 204 abstract CV templates that can capture all the CV VGs found in Hindi. These abstract templates need to be specified for reasons stated above. That is, the V2s in them must be marked so that they are identified by the word variables and not the tag variables. For example for the abstract CV template of V1+V2_fut, the V2 has to be specified with one of the V2s we have listed. For all the 37 abstract universal CV VG templates, we will have 61 templates for each (the total number templates reaching 61*37=2257). For the rest of 167 CV templates, we identify the specific V2s that can play a part in them and draw the actual templates accordingly.

6 Verb Group Tags

Tags/Labels are assigned to the verb groups once identified, at the end of their boundary. These tags are meant to identify the TAM of the verb group. Additionally, these tags also identify the voice form of the verb groups and whether it contains a compound verb construction or not. The general construction of the TAM tags is as follows:

VG.Tense.Aspect.Mood

Here the dot '.' functions as the delimiter between two categories. The first label is 'VG' which identifies that the chunk is a verb group. The second label delimited by the '.' indicates tense of the verb group. Similarly, the third and the fourth indicate aspect and mood. The second label of tense can get prefixed by 'pas' or 'cv' or 'cv_pas'. When prefixed with 'pas', the TAM tag indicates that the VG is in passive voice and when prefixed by 'cv', the VG indicates that it contains a compound verb construction. If it is prefixed by 'cv_pas', it indicates that it is in passive voice and also contains a compound verb construction.

7 Evaluation

We developed a tool[1] to implement this mechanism and ran it over the LDC corpus annotated in the IL-PoST framework. The results obtained in the first run over the LDC corpus were not very promising. We had at least one error in a total of 46% of the sentences. An error analysis showed that these errors were mainly of two types: they contain a verb group which could not be identified and hence marked as 'Unknown' or there are verb groups consisting of no VM and hence the VAUXes are left orphan.

The number of orphan VAUXes was large. These were basically annotation errors in the input corpus itself. Once we corrected them, the tool identified them correctly.

[1] The tool can be tested at the following two sites:
http://sanskrit.jnu.ac.in/vgt/index.html
http://www.langlex.com/vgt/index.html

There can be various reasons for a VG being not recognized. The most incriminating for this test would be the failure of any of the templates to recognize it. Fortunately, that is the case with only a few of the sentences. The most number of unknown occurrences are due to other reasons. The numerous among these is annotation errors. For example there are annotations marked for perfect aspect instead of imperfect and vice versa. Such errors result not only into 'Unknown' marking of the VGs identified but also at times gives incorrect TAM tag. While the VGs marked as 'Unknown' can be identified and the reasons behind this can be ascertained and corrected, the VGs identified incorrectly is hard to find out. We also found a few language specific idiosyncrasies that could not fall into any of the general templates and to capture them we had to prepare a few specific templates just to capture those idiosyncrasies.

We also had to make some modifications in the IL-PoST framework and bring changes accordingly in the LDC corpus for this test. We made changes in three slots of aspect, mood and finiteness of the tags given to a verb.

In the aspect, we removed the progressive aspect marking label of 'prg'. The 'prg' tag in the LDC corpus is given to perfect form of the auxiliary verb 'rəh' because it is this form that indicates the progressive/continuative and durative aspect marking. But because they are marked as 'prg' in the aspect slot, their information about their being in the perfect form is lost. Besides this aspect is marked only periphrastically and should be left to be understood at the local word grouping (LWG) level as we are doing here.

In the modal slot, we removed the marking for habitual mood. Habitual mood marking in Hindi is concurrent with the imperfect marking aspect i.e. both gets realized only if a verb form with the /-tV/ ending form of the verb is present in the verb group. Habituality is understood only syntactically. Instead of habitual mood marking, we marked the /-tV/ ending verb forms as imperfect 'impf'. Another mood, optative, marked morphologically in Hindi is totally missing from the IL-PoST annotation guidelines[2] for Hindi. So we included this mood with the tag of 'opt' in the tags and marked them wherever required.

Under the finiteness slot, digressing from the current guidelines, we suggest that it is verb roots themselves that should be marked as non-finite 'nfn' and all the '-nV' ending verbs be marked as infinitive with the tag of 'ifn'.

After making these changes in the LDC corpus, we ran the tool again over this corpus and this time 99% of the sentences did not show up any issues in identification. For the ones that showed up an issue (a total of 45 out of 4832 sentences), a minor change in the code to capture the tags of the auxiliary verbs can fix them. And thus we can achieve a hundred percent accuracy in identifying the verb groups in Hindi and thereby other major Indian languages having similar structures.

8 Conclusions

In this paper we have shown that identification of the verb groups in Hindi and their TAM can be achieved through a rule based method. We have also shown that through this method we can ascertain whether the VG is in passive voice and whether it also contains a compound verb construction. At present, wherever this task is required, it is done through statistical methods which require a lot of data input. There is still a dearth of large annotated corpus in Hindi. In this situation, this rule based method can be of great help in various NLP tasks including machine translation and grammar checking.

Given that the major Indian languages (Indo-Aryan and Dravidian) share a great lot of similarity in their verb group structure, this rule based method can be replicated for these languages as well.

[2] The details of the guidelines along with the LDC corpus is available at the LDC website: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2010T24

References

Abbi, Anvita, (1992). *The Explicator Compound Verbs: Some Definitional Issues and Criteria for Identification*. In: Indian Linguistics. Vol. 51. No.1.

Aikhenvald, Alexandra and R.M.W. Dixon. (eds). (2006). *Serial Verb Constructions: A Cross-Linguistic Typology*. Oxford University Press.

Bali, Kalika, Monojit Choudhury, Priyanka Biswas, Girish Nath Jha, Narayan Choudhary & Maansi Sharma. (2010). *Indian Language Part-of-Speech Tagset: Hindi*. Linguistic Data Consortium, Philadelphia. ISBN: 1-58563-571-5

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 International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Guru, Kamataprasad. 1978. Hindi Vyaakaran. Nagari Pracharini Sabha: Kashi.

Kachru, Yamuna. (2006). Hindi. John Benjamins: Amsterdam/Philadelphia.

Hook, Peter E. (1974). *The Compund Verb in Hindi*. The Michigan Series in South and South East Asian Languages and Lingusitics: The University of Michigan.

Nespital, Helmut (1997) *Lokabhaaratii: Hindii kriyaa-kosha*. Lokbharti Prakashan, Allahabad.

Paul, Soma. (2006). *An HPSG Account of Bangla Compound Verbs with LKB Implementation*. Unpublished PhD thesis submitted to University of Hyderabad.