Hindi TimeBank: An ISO-TimeML Annotated Reference Corpus

by

Pranav `Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, Manish Shrivastava

Report No: IIIT/TR/2020/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

Proceedings of the 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16), pages 13–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Hindi TimeBank: An ISO-TimeML Annotated Reference Corpus

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi & Manish Shrivastava
Language Technologies Research Center
International Institute of Information Technology
Hyderabad, India
{pranav.goel, suhan.prabhuk, alok.debnath, priyank.modi}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

ISO-TimeML is an international standard for multilingual event annotation, detection, categorization and linking. In this paper, we present the Hindi TimeBank, an ISO-TimeML annotated reference corpus for the detection and classification of events, states and time expressions, and the links between them. Based on contemporary developments in Hindi event recognition, we propose language-independent and language-specific deviations from the ISO-TimeML guidelines, but preserve the schema. These deviations include the inclusion of annotator confidence, and an independent mechanism of identifying and annotating states (such as copulars and existentials) With this paper, we present an open-source corpus, the Hindi TimeBank. The Hindi TimeBank is a 1,000 article dataset, with over 25,000 events, 3,500 states and 2,000 time expressions. We analyze the dataset in detail and provide a class-wise distribution of events, states and time expressions. Our guidelines and dataset are backed by high average inter-annotator agreement scores.
Keywords: Annotation Corpora, Hindi, Temporal Information Extraction

1. Introduction

Temporal information retrieval is a rapidly growing branch of natural language processing and information extraction, due to numerous applications such as question answering and summarization systems. The detection of events, states, temporal expressions and their relations provides a rich source of temporal information, and acts as the representation of real world information in text. This has two-fold implications, first, that the representation mechanism depends on the syntactic and semantic properties of the language, and second, that in order to create systems that use this information, large amounts of annotated data are a prerequisite.

An attempt towards solving the issue of disparate representations was made by ISO-TimeML (Pustejovsky et al., 2010), by developing an international standard based on the earlier, highly popular event annotation framework known as TimeML (Pustejovsky et al., 2003a). ISO-TimeML is an inter-operable semantic framework for linguistic annotation of temporal expressions such as events (e.g. occurrences and happenings) and time expressions (e.g. mentions of days, dates and times). The international standard had been created such that the annotation framework could be applied across languages extensively. The issue of training data for large systems was solved by creating large annotated corpora based on the prevalent annotation mechanism, known as TimeBanks. After the English TimeBank (Pustejovsky et al., 2003b), TimeBanks have been developed for various languages, elaborated upon in section 2..

Recently, Hindi has been added to the list of languages with literature working towards the annotation of events and temporal expressions. Temporal expression identification in Hindi and their basic classification has been done as a part of the FIRE 2011 corpus[1], but a more focused approach has also been adopted (Ramrakhiyani and Majumder, 2013; Ramrakhiyani and Majumder, 2015). For event detection and recognition, the framework and basic guidelines for a binary recognition of event nugget has been established by Goud et al. (2019), which also differentiates between events and states and makes a case for the complexity of recognition of events in a semantico-syntactic grammatical framework of Hindi. We continue to follow this distinction, as mentioned in section 3..

Event analysis as a temporal phenomena is a question not only in NLP but also in linguistic philosophy, which is deeply rooted in the manner languages express events. ISO-TimeML event schema is an improvement over the TimeML event analysis framework to make it more general for all languages. TimeML's definition of an event seems to be derived from the Davidson's notion of eventualities (Davidson, 1967), which provided a definition of events as "*spatio-temporal phenomena with functionally integrated participants*". Therefore, extensionally, TimeML events (Pustejovsky et al., 2007) are based on a neo-Davidsonian analysis of eventualities, and are detected based entirely on properties.

In this paper, we extend both the idea and the initial seed dataset from binary event classification in Hindi (Goud et al., 2019) and include the annotation of states (which was deliberately eliminated earlier), the classification of both events and states, as well as inclusion of time expressions in an augmented dataset of 1,000 Hindi articles. We provide a comprehensive set of guidelines for the identification and differentiation of events from states. The classification scheme for events and states in Hindi has been augmented from TimeML for consistency. Further, changes have been implemented in later stages of the annotation cycle. These are both language specific changes and changes to the ISO-TimeML schema that can be applied to other languages as well. Lastly, the robustness of the annotation guidelines is evaluated by inter-annotator scores, as well as other statistics about the dataset.

To summarize, this paper contributes a corpus of 1,000 articles with 25,829 events and 3,516 states for the purpose

[1] http://fire.irsi.res.in/

of temporal information retrieval in Hindi, the Hindi TimeBank. This resource has been annotated on a modified ISO-TimeML schema and guidelines, which have been elucidated below. We provide a comprehensive analysis of the data, the schema, the guidelines and the annotation mechanism, which can be used for event and temporal expression annotation of multiple other languages.

2. Related Work

TimeBanks have been introduced for multiple languages after English. These TimeBanks were developed after fundamental additions and modifications to ISO-TimeML guidelines for language specific syntactic properties.

In the French TimeBank(Bittar et al., 2011), the authors propose that those verbs be tagged as modal since modality is expressed by fully inflected verbs. Furthermore, the authors also provide a way of capturing the difference between support verb constructions with a neutral aspectual value (mener une attaque (carry out an attack)) and those with an inchoative aspectual value.

The Italian TimeBank(Caselli et al., 2011) focuses on the `EVENT` and `TIMEX3` tag and modifies their properties to suit Italian. The main difference with regards to the `EVENT` tag is in the tag attribute list and attribute values. The `TIMEX` tag used in the Ita-TimeBank is as much as possible compliant with the TIDES TIMEX2[2] annotation.

In the Romanian (Forascu and Tufiș, 2012) and Spanish (Saurı and Badia, 2012) TimeBanks, the authors opted to indicated whether an EVENT is a state (with the 'class' attribute having the value 'STATE'), instead of using the attribute 'type' to indicate if the EVENT is a state, a process or a transition.

The Portuguese TimeBank (Costa and Branco, 2012) uses the same guidelines as the English TimeBank, and use a combination of the Portuguese OpenWordNet and temopral-aware systems.

Finally, in the Persian TimeBank (Yaghoobzadeh et al., 2012), gerund phrases, known as *"esm-e masdar"*, must always be annotated as events, even when they represent generic events. Furthermore, the authors also consider objective deverbal adjectives in PersTimeML. Syntactically, Persian TimeBank differs from ISO-TimeML in the way that all the tokens part of an event are marked under the same event ID irrespective of whether they are consecutive or not.

3. Annotation Guidelines

In this section, we shall cover the basic guidelines for the annotation of events and states, their classification mechanism and the annotation and classification of TIMEX3 time expressions. We present the modified definitions and then use the relevant syntactic cues which will be used in order to determine, annotate and classify both events and states.

3.1. Events and States

TimeML defined *events* as situations that occur, hold or take place, or as states or circumstances in which something obtains or holds true (Saurí et al., 2006). In annotation of Hindi events by this definition, annotators portrayed low confidence, given that event normalization, subordinating verbs and "generics" were not to be marked.

Therefore, states have been defined with respect to these distinctions in order to be easier to annotate. A state may be defined as *a verbal predicate which provides a spatio-temporal description participating entities, including a description of properties, location or existence.* Such a definition accounts for verbal modifiers and copular constructions. Note that subjunctives are not considered states under this definition and neither are they considered events, due to the fact that the participating entities do not undergo any change (Goud et al., 2019). Therefore, subjunctive phrases are not annotated as either events or states.

Furthermore, given an extensional understanding of events based on the change in the properties of entities, certain reporting verbs with sentential predicates are not considered events if they do not contain a participating entity. Hindi allows subject ellipsis constructions, therefore those verbs do not contain any entities, and are therefore not annotated. For example:

1. *kahā jātā hai tūphāna gabhīra hai*
   say go is storm serious is
   It is said that the storm is serious.

Due to the lack of expletive subjects, the verb "*kahA*" can not be attributed to any entity.

3.2. Time Expressions

Time expressions are defined as a span of text which denote a specific time, the duration of an event or state, or a point in time relative to an event or time expression (Group and others, 2009). Annotation and evaluation of temporal annotations is a fundamental concept in information retrieval based on events, as events are anchored on time expressions and therefore it is ubiquitous in semantic evaluation literature (Verhagen et al., 2010).

A time expression consists of a `t_id` which is a unique ID given to each time expression which is useful when they act as anchors to `TLINK`s (explained in Section 3.5.), a `class` which can be a `DATE`, `TIME`, `DURATION` and `SET`, the `tokens` in the span of the time expression and the `AnnConf` (annotator confidence parameter).

The classes of time expressions in Hindi are described as follows:

- `TIME`: The `TIME` category is used to annotate times of the day, which may be specific such as *pAnch baje* (5 o'clock) or a general period such as *subah* (morning). Note that the case markers or *karakas* associated with the time expression are also considered as a part of the time expression when it provides durativity information. For example:

  2. *āja* **pāca** **baje** *vaha* *āegā*
     Today five o'clock he/she come
     He/She will come at 5 o'clock today.

- `DATE`: The `DATE` category is used to annotate calendar days and dates, weekdays and other temporal expressions which consist of multiple days or dates, such

[2]https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf

as weeks, months or years. Note that spans of time with specified start and end dates are not considered in this category. For example:

3. **do** **mahīne** *bāda* *vaha* *āegā*
Two months after he/she come
He/She will come after two months.

- DURATION: The DURATION category is applicable to spans of text which refer to a range of time with start and end times specified in the text.

4. **cāra** **mahīno** *se* *gāyaba* *hai*
four months from missing is
He/she has been missing for four months.

- SET: The SET class of time expressions is used to define the periodicity of an action or refer to an event a definite time in the past or future relative to the current time. The inclusion of *karaka* is important because it denotes the durativity or recursion of the event. For example:

5. **hara** **cāra** **sāla** *olapiksa* *hote* *hai*
every four years Olympics happen is
The Olympics take place every four years.

In using a syntactico-semantic approach to annotate time expressions in Hindi, we need to account for nested time expressions. We do so using a dependency perspective of the time expression itself, by considering the relations between annotations. For example:

6. **bīte** **sāla** **apraila** *se* **jūna** *taka* -
Past year April from June till -
This past year from April to June -

has the standoff annotation:

```
<TIMEX id="t1" class="DURATION"
tokens="1,2", AnnConf="High"/>
<TIMEX id="t2" class="DATE" tokens="3",
AnnConf="High"/>
<TIMEX id="t3" class="DATE" tokens="5",
AnnConf="High"/>
```

3.3. State Categories

TimeML has an event category for STATE and I-STATE. However, as mentioned in Section 4., we do not consider states or intentional states to be events and therefore present the following schema for categorizing states on more syntactic rather than semantic grounds. The category of states introduced in the schema are declarative (DECL) and descriptive (DESC) states.

- DECL: A verb is marked as a declarative state if it provides information about the properties or attributes of a participating entity. They are uniquely identified by copular constructions. For example:

7. *yaha* *gāī* *lāla* *raga* *kī* **hai**
this car red colour (gen.) is
This car is red in colour.

- DESC: A verbal modifier or participle is marked as a descriptive state when it can be rephrased as a copular, and as a modifier provides information about the entity or event it is describing. For example:

8. **khelatā** **huā** *baccā* *pahāī* *karegā*
playing doing child study will do
The child who is now playing will study.

3.4. Event Categories

The event categories are mostly the same as the TimeML event categories (Pustejovsky et al., 2003a). Therefore, the annotated event categories are:

- REP: Reporting events, marked by REP are those events in which an event or state is explained, talked about, spoken, written about or reported. For example:

9. *maine* **kahā** *mujhe* *bhūka* *lagī* *hai*
I said I hunger feel is
I said that I feel hungry.

- ASP: Aspectual events, marked as ASP are the events which denote the beginning, ending, continuation or any other aspectual state of another event. For example:

10. *maine* *khānā* **śurū** **kara** **diyā**
I eating start do did
I started eating.

- PER: Perception events are those events which involve the direct sensing of an event or entity, such as sight, sound or taste. Perception events require an experiencer. For example:

11. *maine* *use* **dekhā** **thā**
I him see had
I had seen him.

- IAC: I_Action events, marked as IAC are the events which explicitly introduce another event as an argument, but not as the aspectual state of that event. In Hindi, there are two syntactic types of IAC events. Either the I_Action occurs as the main verb of the sentence, subordinating the other verb in the sentence, or as the subordinating verb itself. In either case, the I_Action is incomplete without another event or state. For example:

12. *me* **padhakara** *so* *jāūgā*
I read after sleep will go
I will go to sleep after reading.

- OCC: All other events, which are not categorized above are categorized as occurrences, marked as OCC. All nominal events are inherently occurrences. For example:

13. *yuddha* *me* *sainika* **ghāyala** **hue**
war in soldiers hurt got
Soldiers were hurt in the war.

3.5. Linking Events and TIMEX3 Annotations

TimeML introduces three links, known as TLINK, SLINK and ALINK, which are described below (Saurí et al., 2006).

- TLINK: A temporal link or TLINK is a relationship between two events or states (represented by their instance IDs), or of an event or state with a time. It is categorized into BEFORE, BEFORE-OVERLAP and OVERLAP (O'Gorman et al., 2016).

14.

yuddha	*me*	*sainika*	**ghāyala**	**hue**
war	in	soldiers	hurt	got

Soldiers were hurt in the war.

```
<TLINK f_id='e1', s_id='e2',
class='OVERLAP' AnnConf='High'/>
```

- SLINK: A subordination link or SLINK is used to annotate the relations between two events, specifically reporting and other events. We also consider certain intensional events with other events given that the latter event expects or determines the former event. Conditional constructions are annotated as SLINK as well.

15.

rāma	**kahatā**	**hai**	*kī*
ram	says	is	that
yuddha	**gabhīra**	**hai**	
war	serious	is	

Ram says that the war is serious.

had the annotation :

```
<SLINK f_id='e1', s_id='e2'
AnnConf='High'/>
```

- ALINK: An aspectual link or ALINK shows the relation between an aspectual event and its argument event or state. The ALINK tag had 4 classes viz. INITIATION, TERMINATION, CONTINUATION and CONCLUSION inspired by Pustejovsky et al. (2003a).

16.

rāma	*ne*	*khānā*	**śurū**	**kara**
ram	(Erg.)	eat	start	did

Ram started eating.

had the annotation :

```
<ALINK f_id='e1', s_id='e2',
class='INITIATION' AnnConf='High'/>
```

4. Modifications to ISO-TimeML

In this section, we review some of the basic modifications to the ISO-TimeML event annotation schema and guidelines that have been used to annotate the Hindi TimeBank. The modifications are twofold, one which are language independent or cross lingual, and can be applied for creating new TimeBanks in other languages as well as extending the Hindi TimeBank, and the second which are particular to Hindi due to the semantico-syntactic nature of its grammar.

4.1. Language Independent Modifications

Pustejovsky et al. (2008) introduces the <CONFIDENCE> tag in order to provide the notion of a confidence metric to each attribute of each tag. The confidence tag used was in the range of 0 to 1 and was used to determine the annotator's confidence in every attribute annotation. However, this notion was found to be too granular. Given the attributes we annotate in the Hindi TimeBank, considering annotator confidence as an attribute rather than a standoff tag seemed more appropriate.

In our system, the annotator confidence metric is a ternary annotation parameter with values HIGH, MEDIUM and LOW, and is meant to signify how confident an annotator is about an annotation. Thus, we found that the annotator confidence metric is a very useful parameter in determining the clarity of definitions to annotators specifically in event and time expression classification.

The annotator confidence parameter helps in justifying the changes over iterations of guideline development, and also serve to point of ambiguous constructions which rely heavily on context, and/or represented a facet of the grammar that can not be captured by the current guidelines, and may pose a problem for further processes done using this data. One significant point based on which annotator confidence proved pertinent is the removal of subjunctives from event representation.

4.2. Language Specific Modifications

There are a number of modifications made to the ISO-TimeML guidelines which needed to be made due to the discourse structure and the grammatical framework associated with the identification and classification of events in Hindi. These changes include the identification of states, modifying the classification of events due to state categorization and a entity-centric event descriptions.

Identification of States TimeML presents events "as a cover term for situations that happen, occur, hold, or take place as well as those predicates describing states or circumstances in which something obtains or holds true" (Pustejovsky et al., 2003a). However, Goud et al. (2019) mentions the difficulties in direct annotation of events and states from a linguistic philosophy perspective as well as from an annotation guidelines standpoint.

As mentioned in Section 1., TimeML's definition of events seems similar to the syntactically motivated new-Davidsonian definition of an event. However, our analysis of events and states is based on Bach's definition of events, states and processes (Bach, 1986), which is similar to Panini's event semantic representation. We present the need for a separate notion of state by showing the following example from Goud et al. (2019):

17.

ijarāila	*me*	*gaisa*	*māska*	*kī*	***kamī***
Israel	in	gas	mask	of	shortage
se	*unhe*	*takalīpha*	*honī*	*lagī*	
reason	they	hardship	happen	began	

Due to a shortage of gas masks in Israel, they began to suffer.

18.

ijarāila	*me*	*gaisa*	*māska*	*kī*	***kamī***
Israel	in	gas	mask	of	shortage
hone	*se*	*unhe*	*takalīpha*	*honī*	*lagī*
to-be	reason	they	hardship	happen	began

Due to a shortage of gas masks in Israel, they began to suffer.

While these sentences are semantically equivalent, the syntactic representation of the subordinate verb clause is very different, as the presence of the verbal auxiliary *hone* explicitly marks a notion of a telic and adurative situation. According to TimeML, generics and verbal clauses with generic arguments are not to be annotated as events. However, the auxiliary *hone* is used with generics to construct semantically equivalent sentences. Therefore, according to TimeML annotation guidelines, annotators would not mark *kamI* as an event, but would mark *kamI hone*, even though they have been used in the same way in the sentence.

In order to resolve this discrepancy, we turn to the Paninian grammatical framework. The presence of auxiliaries in the verbal predicate are used to denote emphasis, tense and aspect information (Palmer et al., 2009). From the perspective of event and state representation, the auxiliaries are representative of the telic and durative properties of the predicate, which makes both their representation as well as participation of entities different depending on the type of verbal auxiliary used. Therefore:

- Verbal auxiliaries provide syntactic as well as semantic information about the verbal predicate, which is crucial.
- A verbal predicate may therefore be considered either a state or an event if compared to Bach's notion of eventualities

Since Bach's definition helps in the identification and classification of generics, habitual verbal predicates as well as other semantically equivalent but syntactically distinct forms, we adopt its definition for identifying states as a unique concept. Therefore, for the example above, we uniformly mark both *kamI* and *kamI hone* as descriptive states which have been described in the annotation guidelines (Section 3.3.).

We found that on introducing and defining states, annotator confidence regarding verbal modifiers as well as clauses with ambiguous constructions rose significantly, as it made the guidelines more naturally aligned to the annotator's understanding of the language. This solidified the inclusion of states into the Hindi TimeBank.

Modification in Classification Mechanism Given that both states and events are being annotated as independent concepts, the classification prescribed by TimeML (Pustejovsky et al., 2003a) can not be used directly. Instead, the `STATE` and `I-STATE` event categories have been removed. We have seen that our analysis of states are different from the TimeML representation of states. TimeML defines `I-STATE` as states that refer to alternate or possible worlds. Hindi only presents these constructions as subjunctives, which explicitly do not include a participant. Therefore, by definition, `I-STATE` are not annotated as states, but are identified as `OCC` events.

We introduce a classification schema for states, which are `DESC` and `DECL`, the description of which are given in Section 3. For example:

19.

khelatā	**huā**	*baccā*	**bhāga**	**rahā**	**hai**
play	doing	child	run	-ing	is

The child who is running, is playing.

has the standoff annotations:

```
<STATE id="s1" class="DESC" tokens="1,2"
annconf="HIGH" />
<EVENT id="e1" class="OCC" tokens="4,5,6"
annconf="HIGH" />
```

using the ISO-TimeML XML schema. The phrase "*KelawA huA baccA*" translates to "the child who is playing", but the verb form used is a verbal modifier and not a participle, and therefore it does not change the state of the participating entity, rather describes it.

Modifications to TIMEX3 TIMEX3 in Hindi has been studied (Ramrakhiyani and Majumder, 2013; Ramrakhiyani and Majumder, 2015) in order to analyze, identify and classify time expressions in Hindi. The procedure for annotating and extracting time expressions manually has been detailed in section 3.. We deviate in the annotation of fragmented time expressions by taking only those tokens which give us a local time expression and grouping them under a single `TIMEX` id. Relative time expressions such as *cāra sāla* (four years) can only be annotated as a `TIMEX` only if the duration can be estimated. We also account for dependency and semantic role information when annotating time expressions, which is not considered in TIMEX3.

5. Annotation Pipeline

Goud et al. (2019) proposed an event tagged dataset comprising of 810 news articles, which were primarily from the financial and crime domain, annotated only by the presence of events. We discarded all the articles which had less than 100 tokens, since these files did not contribute to the information base.

We chose a group of 8 annotators for the task of annotation as well as evaluation of the bootstrapped dataset. The annotators are native Hindi speakers, educated in both English and Hindi. All annotations were carried out using the BRAT Annotation Framework (Stenetorp et al., 2012). Figure 1 shows the annotation procedure in detail. There are multiple rounds of annotation in each stage of the pipeline.

5.1. Event and State Identification

Since Goud et al. (2019) only had events identified, the first task at hand was to annotate the states in these articles.The files were annotated by 8 annotators in batches of 100 articles, over 2 rounds of annotation.

Since all the articles from Goud et al. (2019) were from the financial crime domain, this dataset was not balanced well in terms of the types of syntactic and semantic environments in which events and states can occur in Hindi.

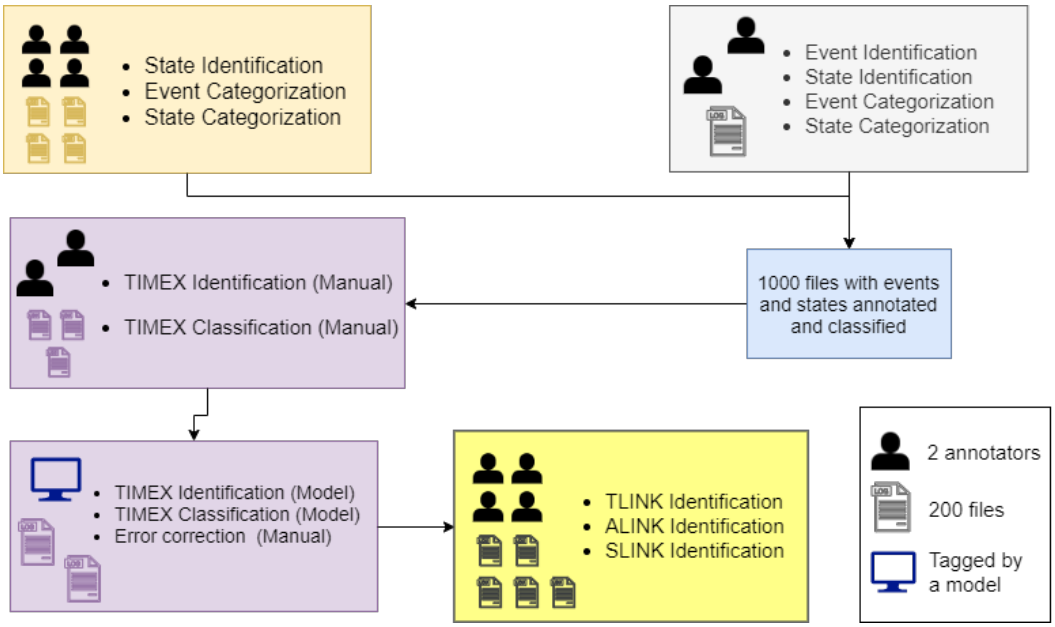

Figure 1: Annotation Steps for Hindi TimeBank. The legend for each icon used in the diagram is provided in the bottom right.

Thus, 200 articles were added to the seed dataset, out of which 150 were news articles and the remaining 50 were short fiction stories. We collected these news articles from Navbharat Times[3], a national Hindi daily newspaper with over 2 million copies circulated nationwide. The distribution of these scraped articles can be found in Table 2. The short stories are by Premchand, who is a renowned Hindi author[4]. The addition of these articles will allow the models trained on the Hindi TimeBank to be more reliable in detecting events, states, and temporal expressions in Hindi text.

For these 200 articles, they were first tokenized using a freely available tokenizer (Reddy and Sharoff, 2011) [5] and then the identification of both events and states were done by 4 annotators in batches of 50 articles over 4 rounds. Large inter-annotator disparity was found between annotators for reporting verbs with no participating entity, due to which those constructions were removed from the purview of event and state annotation.

5.2. Event and State Categorization

The above mentioned 200 articles were annotated for event and state categories by 4 annotators in batches of 50 articles over 4 rounds. The classification guidelines are based on easily identifiable syntactic differences, which made the manual annotation of events and state categories a high-confidence task among annotators. Once these 200 articles were annotated, the dataset of Goud et al. (2019) was annotated for the same by 8 annotators in batches of 100 articles over 3 rounds of annotation.

[3] https://navbharattimes.indiatimes.com/

[4] https://hindisamay.com/premchand%20samagra/Indexpremchand.htm

[5] https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger/src/master/

Features	Description
WI	Word Identity
POS	Part-of-Speech
BT	Bi-gram and tri-gram features
BOS	Beginning Of Sentence
ISTIMEX	Current Word is part of a TIMEX tag

Table 1: CRF Features

This resulted in a corpus of 1000 articles with event and state phrase boundaries identified and classified.

5.3. TIMEX Annotation and Classification

Automated Identification: For the first sub-task, our CRF model was trained on the set of 600 articles tagged manually and tested on the remaining 400 articles, in which, time expressions were identified. This CRF used the first 4 features of Table 1 and had a precision of 0.79 in this sub-task. The resultant labeling was evaluated manually by 4 annotators, and the relevant changes to the dataset were made.

Automated Categorisation: For the second sub-task, which was the categorization of the annotated time expressions, our CRF was trained on the set of 600 articles tagged manually and tested on the remaining 400 articles. For this CRF, the `ISTIMEX` feature of Table 1 was used in addition to the rest of the features. Our CRF had a precision of 0.84 in this sub-task. The labeled data was then corrected manually by 4 annotators in 2 rounds of annotation.

Finally, the resultant dataset was manually annotated with temporal links (`TLINK`), aspectual links (`ALINK`) and and subordination links (`SLINK`). This phase of annotation required 8 annotators with 4 rounds of annotations in batches of 125 articles each.

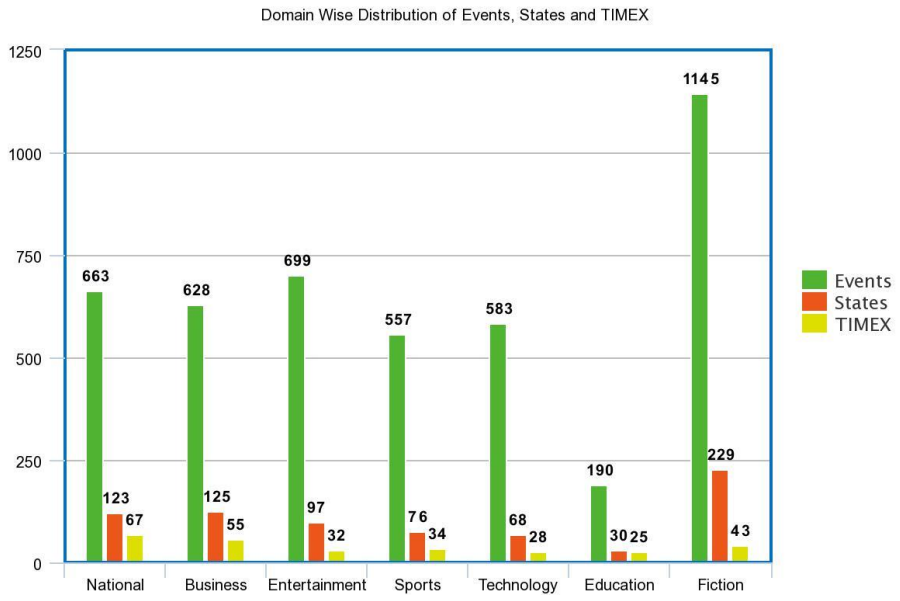

Figure 2: Domain Wise Distribution for Event, State and TIMEX tags

Domain	Number of Articles
Financial (Goud et al., 2019)	800
Fiction	50
National (News)	30
Business Analysis (News)	30
Entertainment (News)	30
Sports (News)	25
Technology & Development	25
Education (News)	10
Total	1000

Table 2: Distribution of Articles by Domain

Feature	Total
Number of Tokens	292,517
Number of Events	25,829
Number of States	3,516
Number of TIMEX	2,396
Number of TLINK	7,289
Number of SLINK	4,741
Number of ALINK	433

Table 3: Count of Event, States, TIMEX and all types of links

6. Corpus Statistics

In this section we present some basic statistics of the Hindi TimeBank, such as the number of events, states, categories and links. As mentioned in the Section 5., we annotate 1000 articles from multiple domains. Table 3 shows the total number of events, states, TIMEX and all of the links in the corpus. In the following subsections, we present the ratio of classes of events, states, time expressions and links. We also present the statistics on annotator confidence and the inter-annotator agreement scores.

	Category	Total
Event	OCC	22,606
	REP	1,599
	IAC	783
	ASP	421
	PER	420
	Total	25,829
State	DESC	1,865
	DECL	1,651
	Total	3,516

Table 4: Event and State Categories and Distribution

6.1. Event and State Statistics

In this section we provide insight into the distribution of the event and state categories. Table 4 provides the details of the distribution of events in the dataset. We see that the occurrence type (OCC) is the most popular, accounting for 87.52% of the total number of events. The aspectual type (ASP) accounts for 1.62%, the intensional action (IAC) for 3.03%, the perception events (PER) for 1.62%, and the reporting (REP) event 6.19% of the total events.

The occurrence type is the most popular type of event due to limited syntactic and semantic constraints on its classification and the fact that an event was annotated as an occurrence if did not belong to any other category.

We provide a similar analysis of states, with 53.04% of the states being descriptive (DESC) and 46.96% being declarative (DECL) in nature.

In Figure 2, we show the domain wise distribution of Event, State and TIMEX tags. We observe that the number of events are significantly higher than the number of states and time expressions across all domains. For Goud et al. (2019), the number of events, states and time expressions are 21,364, 2,768, and 2,112. These numbers are not represented in Figure 2 as they account for 800 articles (80%) of the dataset.

TIMEX Category	Total
DATE	1,390
DUR	545
TIME	433
SET	28
Total	2,396

Table 5: Time Expressions Categories and Distribution

6.2. Time Expression Statistics

In this section, we look into the time expressions in the Hindi TimeBank. We see in Table 5 that a majority of the time expressions belong to the $DATE$ class.

6.3. Annotator Confidence and Inter-Annotator Agreement Scores

In this section, we calculate the inter-annotator agreement scores for the event and state detection. This is done by Fleiss' Kappa metric (Fleiss and Cohen, 1973) as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where $P - P_e$ is the actual degree of agreement achieved and $1 - P_e$ is the degree of agreement above chance. Given N tokens to be annotated and n annotators, with k categories to annotate the data. We first calculate the proportion of annotations in the j^{th} domain as:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \ , \quad 1 = \sum_{j=1}^{k} p_j \tag{2}$$

We then calculate P_i, the degree of agreement with the i^{th} annotator as:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{3}$$

$$= \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^{k} n_{ij}^2 \right) - n \right] \tag{4}$$

Finally we calculate $\bar{P}$ and $\bar{P}_e$ as:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{5}$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{6}$$

We also provide the domain wise breakdown of annotator confidence in the final corpus in Table 7. We do not remove annotations which are marked as MEDIUM or LOW by the annotators. Annotator confidence variations are seen most for events which have some ambiguity of being considered states. Lower confidence is associated with those verbal predicates which have only tense auxiliaries but either belong to a fragmented event or are in light verb constructions. State annotations show low confidence for descriptive states which are emphasized. TIMEX classification has no low confidence scores. Classification causes some low and medium confidence scores among TLINKS and ALINKs. In the case of SLINKs, subordination of OCC–OCC links are most ambiguous and result in low confidence among the annotators.

Annotation	Fleiss' Kappa Score
Detection of Events	0.84
Detection of States	0.81
Event Categories	0.77
State Categories	0.86
TIMEX Detection	0.88
TIMEX Categories	0.86

Table 6: Inter Annotator Agreement for Various Annotation Phases

Category	High	Medium	Low
Event Categories	92.24%	5.86%	1.90%
State Categories	91.07%	5.52%	3.41%
TIMEX Categories	95.69%	4.31%	0.00%
TLINK	90.86%	4.25%	4.89%
ALINK	93.35%	4.60%	2.05%
SLINK	89.77%	5.71%	4.52%

Table 7: Category-wise Breakdown of Annotator Confidence Scores

In the future, we hope this effort can help in the development of TimeBanks for other languages. The current corpus can also be enriched with the annotation of relations between the events and states based on causality and correlation. In it current form, the corpus can be used for generating a minimal knowledge graph which may also be enriched by entity and event linking.

7. Conclusion

In this paper, we present the Hindi TimeBank, a large event, state and time expression annotated corpus. We describe the annotation mechanism and modifications we made to the ISO-TimeML guidelines in order to annotate the data. We provide extensive analysis of the annotation methodology, so that the process of creating TimeBanks for other languages can be a structured effort, especially for languages with similar syntactic and semantic constraints as Hindi. We also present a detailed analysis of the corpus itself, including the distribution of events and states, their categories and the links between them, as well as the distribution of extents and types of time expressions.

The Hindi TimeBank has been created such that it can be used to further event annotation and detection research in Hindi, and the modifications to ISO-TimeML can be used to annotate TimeBanks for other Indo-Aryan languages. The current corpus can also be enriched with the annotation of relations between the events and states based on causality and correlation. A better annotation by nuanced clustering of dates as a duration, and the analysis of TIMEX types such as duration and set is also a direction for further exploration. In its current form, the corpus can be used for generating a minimal knowledge graph, which may also be enriched by entity and event linking. The corpus can also act as a gold standard dataset for machine learning applica-

tions for Hindi.

To get access to the dataset, please e-mail the authors [6].

8. Bibliographical References

Bach, E. (1986). The algebra of events. *Linguistics and philosophy*, 9(1):5–16.

Bittar, A., Amsili, P., Denis, P., and Danlos, L. (2011). French timebank: an iso-timeml annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 130–134. Association for Computational Linguistics.

Caselli, T., Bartalesi Lenzi, V., Sprugnoli, R., Pianta, E., and Prodanof, I. (2011). Annotating events, temporal expressions and relations in Italian: the it-timeml experience for the ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Portland, Oregon, USA, June. Association for Computational Linguistics.

Costa, F. and Branco, A. (2012). Timebankpt: A timeml annotated corpus of portuguese. In *LREC*, pages 3727–3734.

Davidson, D. (1967). The logical form of action sentences.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Forascu, C. and Tufiș, D. (2012). Romanian timebank: An annotated parallel corpus for temporal information. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Goud, J. S., Goel, P., Debnath, A., Prabhu, S., and Shrivastava, M. (2019). A semantico-syntactic approach to event-mention detection and extraction in hindi. In *Workshop on Interoperable Semantic Annotation (ISA-15)*, page 63.

Group, T. W. et al. (2009). Guidelines for temporal expression annotation for english for tempeval 2010.

O'Gorman, T., Wright-Bettner, K., and Palmer, M. (2016). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Palmer, M., Bhatt, R., Narasimhan, B., Rambow, O., Sharma, D. M., and Xia, F. (2009). Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Pustejovsky, J., Littman, J., and Saurí, R. (2007). Arguments in timeml: events and entities. In *Annotating, Extracting and Reasoning about Time and Events*, pages 107–126. Springer.

Pustejovsky, J., Lee, K., Harry, H. B., Boguraev, B., and Ide, N. (2008). Language resource management—semantic annotation framework (semaf)—part 1: Time and events. *International Organization*.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Ramrakhiyani, N. and Majumder, P. (2013). Temporal expression recognition in hindi. In *Mining Intelligence and Knowledge Exploration*, pages 740–750. Springer.

Ramrakhiyani, N. and Majumder, P. (2015). Approaches to temporal expression recognition in hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 14(1):2.

Reddy, S. and Sharoff, S. (2011). Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 11–19, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Saurı, R. and Badia, T. (2012). Spanish timebank 1.0. *LDC catalog ref. LDC2012T12*.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). Timeml annotation guidelines. *Version*, 1(1):31.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.

Yaghoobzadeh, Y., Ghassem-Sani, G., Mirroshandel, S. A., and Eshaghzadeh, M. (2012). Iso-timeml event extraction in persian text. In *Proceedings of COLING 2012*, pages 2931–2944.

[6] mailto: pranavgoel25[at]gmail.com