A Semantico-Syntactic Approach to Event-Mention Detection and Extraction In Hindi

by

Jaipal Singh Goud, Pranav `Goel, Alok Debnath, Suhan Prabhu, Manish Shrivastava

in

13th International Conference on Computational Semantics (IWCS 2019)
(*IWCS-2019*)

Report No: IIIT/TR/2019/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2019

A Semantico-Syntactic Approach to Event-Mention Detection and Extraction In Hindi

Jaipal Singh Goud
IIIT Hyderabad
jaipal.singh@research.iiit.ac.in

Pranav Goel
IIIT Hyderabad
pranav.goel@research.iiit.ac.in

Alok Debnath
IIIT Hyderabad
alok.debnath@research.iiit.ac.in

Suhan Prabhu
IIIT Hyderabad
suhan.prabhuk@research.iiit.ac.in

Manish Shrivastava
IIIT Hyderabad
m.shrivastava@iiit.ac.in

Abstract

This paper introduces a gold standard event-annotated dataset of 810 Hindi news article as well as a set of comprehensive guidelines for detecting and annotating events in Hindi. We present our linguistically motivated guideline development process, with a focus on annotator friendliness, that can be replicated for event mention detection for most Indo-Aryan languages. The paper highlights the challenges of detecting event mentions in Hindi given the unique semantic constraints on the syntactic apparatus used for denoting events. Our work as a whole also establishes a language agnostic pipeline for the development of an event annotated corpora and event detection guidelines.

1 Introduction

Event detection is an important problem in both natural language processing and information retrieval. It has been extended to multiple domains and is now being tackled for languages other than English. In this paper, we address the problem of event detection in Hindi. This task is relatively unexplored for Hindi, partly because of syntactic features such as relative free word order and nature of verb adjuncts and partly due to the lack of structured data. The syntax of the language increases the number of options available for expressing structured information, hence significantly complicating how chunks of text can be semantically interpreted. Therefore, our approach to event detection is based on defining events in a semantico-syntactic idea, meaning that certain clause, phase and sentence structures are *eventive* in nature.

This task is an important one given the constant rise in volume of Hindi data being produced and consumed. A recent report shows that Hindi content consumption on the web is growing at a rate of 94% per year as compared to English's 19% [1]. From the perspective of extracting latent information, defining events as semantico-syntactic objects allows for extracting relations between both events and entities, that are not explicitly mentioned. The extracted event information can then be readily consumed by downstream tasks such as question answering, temporal and causal inference detection and summarization.

[1] https://economictimes.indiatimes.com/tech/internet/hindi-content-consumption-on-internet-growing-at-94-google/articleshow/48528347.cms

Therefore, in this paper, we first define events in Hindi and explore the challenges involved in detection of these event in Hindi news text. We then define a schema and a set of guidelines for event detection in Hindi. Here, the schema refers to the various components of the annotated textual span, such as the event trigger, nugget and core, while the guidelines define the mechanism for the annotation of events based on syntactic characteristics of the schema. The guidelines, while inspired by TimeML (Pustejovsky et al., 2003) in format and manner, are **not** a direct adaptation of the TimeML guidelines, as that involves event classification and event relations, which is beyond the scope of this problem. This task is more geared towards establishing events as a concept that can be annotated in Hindi. We tackle the challenge of event detection from the perspective of maximum capture of event information given the nuances of Hindi grammar.

This task is daunting for a language like Hindi, because its syntactic properties (such as *karakas* and verb complements) tend to be governed by the semantics of the sentence, which is further emphasized by the relatively free word order. Therefore, the combination of our schema and guidelines explains the possible sentential structures in which an event can occur, accounting for characteristics such as fragmentation, stative constructions and negation.

2 Related Work

Automated event detection as a problem has been studied for a long time. A fundamental issue in solving this problem lies in how we define events in a given context (Goyal et al., 2013). To the best of our knowledge, there has been no prior work on a semantico-syntactic definition of events in Hindi. There have been multiple attempts to define events in English, but due to the varied complexities in definitions of events, most of the research focuses on limited guidelines for event detection in English texts like ACE (Doddington et al., 2004), TimeML (Pustejovsky et al., 2003), RED (O'Gorman et al., 2016), ECB (Bejan and Harabagiu, 2010), and ERE (Song et al., 2015). TimeML, a specification language for events and temporal expressions occurring in English texts, was developed to enhance natural language question answering systems about events and entities in news articles as a part of the TARSQI (Temporal Awareness and Reasoning systems for QA) Project[2]. TimeML guidelines have been adapted to languages such as Korean (Im et al., 2009), Italian (**?**), French (Bittar et al., 2011a) and Persian (Yaghoobzadeh et al., 2012). Hindi event detection was motivated and inspired by TimeML (Pustejovsky et al., 2003),

All the above mentioned guidelines were developed in order to structure the data for a particular system, context or purpose. ECB was developed for identifying coreference in news articles; ACE for the purpose of automatic inference of entities, relations and events from data. ERE was developed as a lighter-weight version of ACE with the goal of making annotation more convenient by consolidating some of the annotation type distinctions that were found to be the most problematic in ACE, as well as removing some more complex annotation features, and making the annotation more consistent across annotators.

ISO-TimeML (Pustejovsky et al., 2010) presents an international standard markup language for the annotation of events using the <EVENT> tag, temporal expressions using the <TIMEX3> tag, and links between entities such as temporal links <TLINK>, aspectual links <ALINK> and modal subordination <SLINK>. Multiple languages have datasets annotated based on ISO-TimeML including French (Bittar et al., 2011b) and Italian (Caselli et al., 2011). Our work deviates from ISO-TimeML in that we only define events for Hindi in this paper. The motivation for this deviation is that the definition of events in both TimeML and ISO-TimeML are syntactic in nature, whereas our definitions have a semantic basis, leading to a change in the very nature of the annotation.

[2]http://www.timeml.org/tarsqi/index.html

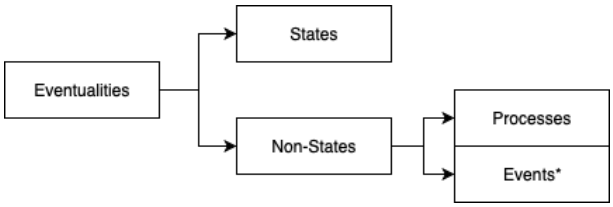

Figure 1: Bach's Eventualities

3 Introduction to Events and States

Defining the concept of events is deeply rooted in the linguistic study of verbs. In this section we lay the foundation for our definition of events. We also highlight the differences between states and non-states and propose our definition of events.

3.1 Eventualities

Initial works of Vendler (1957) on aspectual classification of verbs, classify verbs into four broad categories: states, accomplishments, achievements and activities, but it was the work of Bach (1986) that brought forth the notion of **eventualities**. He defined eventualities as aspectual phenomena and proposed that they were a "broader" notion of events. The idea was to include all durative phenomena which are telic and atelic, such that they could be classified as states, processes and events. The three components are defined using concepts of durativity (having a duration) and telicity (having an explicit condition of termination) (Moens and Steedman, 1988). Bach (1981, 1986) proposes the following division of 'eventualities'

1. **States** : Notions that are durative and changeless, e.g, hate, resemble
2. **Processes** : Notions that are durative and atelic, e.g, raining, sleeping
3. **Events** [3] : Notions that are momentaneous or telic, e.g, walk, slap, tap

3.2 States and Non-States

The categorization based on durativity and telicity captures the basic distinction between states and non-states. This distinction was de-emphasized in earlier works such that those of Vendler (Vendler, 1957). Bach's eventualities distinguish states from non-states (processes and actions) by recognizing the fact that there is no change involved in states and there is no explicit condition of termination.

Pustejovsky (1991) also followed this taxonomy of states, events and transitions, however, providing a more detailed understanding of subeventual structures in event types. This structure can only exist for non-states, given that any nesting in stative constructions will collapse to give a single description of the entity.

3.3 Events

Following the concept of eventualities as defined by Bach (Bach, 1986), we propose the definition of events as: **eventualities that are non-states and involve zero or more participants and attributes**. Essentially, events in Hindi are a combination of Bach's events and processes, and we exclude states from this definition entirely. This has been done for reasons explained in Section 3.4

Events are independent abstract concepts in the real world and not textual expressions bound to specific words or phrases. Events are lexicalized by languages as event mentions (Section 4.1). Given

[3]This refer's to Bach's distinction of eventualities into events. Our definition of events is introduced in Section 3.3

the unique syntactic constructions and semantic coverage associated with lexical items in different languages, each language captures event mentions differently. Events and event mentions can be better understood by looking at the example below :

1. *rameS* *ne* *KAnA* ***KAyA***
   Ramesh (erg.) food ate
   "Ramesh ate the food."

1. *usne* *apnA* *kAm* ***kIyA***
   He/She (gen.) work did
   "He/She did his/her work."

Here, the event is the notion of eating and doing respectively. The phrases *KAyA* and *kIyA* are **event mentions** that refers to the actual event of eating and doing respectively.

3.4 Deconstructing TimeML Events

Pustejovsky et al. (2003)'s definition of events, in TimeML, can be deconstructed into the following two distinct concepts.

- *We use event as a cover term for situations that happen, occur, hold, or take place* : These are captured as dynamic events (referred to henceforth as states), such as causing or intending to cause a change in the condition of involved entities.
- *We also consider as events those predicates describing states or circumstances in which something obtains or holds true*: These are captured as states, which define the condition of the involved entity at a point of time.

The second part of the TimeML event definition, concerned with states, causes trouble when applied to Hindi. In Hindi, addition of a verbalizer transforms almost all adjectives and nouns into events, however it is possible that the noun or adjective (without the verbalizer) does not actually represent an event. Consider the following example :

1. *Israel* *mein* *gas* *mask* *kI* ***kamI*** *ke* *kAran* *unhe* *AyAt* *baDhAni* *padi*
   Israel in gas mask of **shortage** (gen.) reason they import increase had-to
   Due to a shortage of gas masks in Israel, they had to increase imports.

2. *Israel* *mein* *gas* *mask* *kI* ***kamI*** ***hone*** *ke* *kAran* *unhe* *AyAt* *baDhAni* *padi*
   Israel in gas mask of **shortage** **to-be** (gen.) reason they import increase had-to
   Due to a shortage of gas masks in Israel, they had to increase imports.

In the first sentence, "*kamI*" denotes the state of shortage. While in the second sentence, we see that the verbalizer "*hone*" very distinctly makes the span "*kamI hone*" an event. However, semantically, the statements are equivalent, meaning that the existence of an event in the second implies the existence of an event in the first. Intuitively, and as observed while developing the annotation specifications, it was found that the annotation of states was challenging, due to their lack of syntactic uniformity. This made it difficult for manual annotators to confidently annotate states. To find the annotators' belief in the correctness in their annotation of an event, the *annotator confidence* parameter was introduced. If annotators felt uncertain about an annotation they could choose to give it a low confidence tag. By default every annotation is given a high confidence tag. It was found that states comprised an overwhelming majority of low confidence scores, making it challenging to use these annotations due to the high degree of uncertainty. Hence, for the remainder of this paper, the annotated data is focused *solely on events*.

4 Annotation Schema

In this section, we define the annotation schema for capturing the maximal amount of information in order to capture those textual mentions which are events. We define the terms event mention, event trigger and event core.

4.1 Event Mention

An event mention is the textual span expected to provide complete information of the event in terms of meaning, temporality and aspect. The event mention is therefore composed of four major parts, the event nugget (section 4.2), the tense marker, the aspect marker, and an optional polarity (negation) marker (section 4.4).

4.2 Event Nugget

Mitamura et al. (2015) defines the event nugget as "a semantically meaningful unit that expresses the event in a sentence. An event nugget can be either a single word (main verb, noun, adjective, adverb) or a continuous or discontinuous multi-word phrase."

Single word event nugget: That word in the event mention which contains the most semantic information about the event, and governs its argument structure is the event nugget. Semantically, it governs both the nature and the change in state of the participating entities.

2.

rAkeS	*uskA*	*KAnA*	[***KA***	*rahaa*	*hai*]
Rakesh	his/her	food	eat	(prog.)	is

"Rakesh is eating his/her food"

In the example above, the word ***KA*** is the event nugget, because it is the only word in the event mention *KA rahA hai* that defines its argument structure. The nature of the event is governed by the meaning of *KA*.[4]

Multi-word event nugget: Here the event nugget is a multi-word phrase, that is usually non-compositional in meaning. In case of multi-word event nuggets, the semantics of the event is provided by all the words in the event nugget, therefore the nature of the participating entities are governed by the meaning of the phrase and therefore the change in state of the participants will also be different. For example:

3.

rAkeS	*uskA*	[sar	***KA***	*rahA*	*hai*]
Rakesh	his/her	head	eat	(prog.)	is

"Rakesh is annoying him/her."

The phrase *sar KAnA* literally means to annoy, which takes a set of arguments which are distinct from the nature and change in state of the entities from the event *KAnA* (to eat).

4.3 Fragmented Events

Hindi is a relatively free word order language. A consequence of that is the fragmentation of events which splits the event mention across a sentence, as a result of which the insertion of a phrase (or clause) within the event mention is not ungrammatical. An example of a fragmented event:

4.

rAm	*ne*	***nirdeS***	*gusse*	*se*	***diyA***
Ram	(erg.)	order	anger	with	gave

"Ram gave the order angrily."

[4]The event nugget determines that among the participating entities, one has the ability to eat and the other, to be eaten. This borrows from the Paninian framework of *yogyatA* or semantic expectation (Bharati and Sangal, 1993).

4.4 Capturing Negation

Events with a negative polarity in Hindi are temporally bound. They represent an instance where an event does not occur for a period of time. The positive event can be expected to occur otherwise. The concept of temporal bound of events with a negative polarity is introduced by explicit negative markers such as *nahI*, *nA* and a few more. To account for events with a negative polarity, we mark polarity indicators in the event mention.

5. *rAm* *ne* *KAnA* ***nahI*** ***KAyA***
   Ram (erg.) food not ate
   "Ram didn't eat the food."

5 Annotation Guidelines

While detecting events in Hindi, there are various syntactic features that need to be accounted for. These syntactic considerations are only made while annotating the span of an event. Eventiveness of a textual unit is independent of the syntax. As Hindi is mostly an analytic language, verbs in Hindi have tense and aspect markers as individual lexemes. This, along with the presence of phrasal conjuncts and conjunct verbs (Begum et al., 2011a), results in ambiguity in detection of event mentions as shown below.

5.1 Verbs and Verb Complexes

Verbs are parts of speech that denote action. Therefore, by definition, verbs and verbal predicates are events. Verbs in Hindi can be simple verbs, compound verbs, conjunct verbs and phrasal verbs. Unlike most languages, Hindi has the property of verbal constructions that span multiple words. This is due to the analytical nature of the language, and leads to constructions which can be modeled both semantically or syntactically.

Verbs in Hindi have distinct tense and aspect markers. Furthermore, nouns are verbalized by using a verbalizer, a separate lexical item which is independently a verb, but which in the presence of a noun transforms it into a verb with the same semantic implications as the noun. The list of verbalizers is a closed class of words, but allowing such a construction leads to multiple idiomatic and phrasal constructions.

The syntactic classification of these phenomena provides two basic classifications of verbs: predicate constructions and light verb constructions. However, these classifications do not show the relation of the noun to the nature of the verb, and are therefore not enough to explore the eventive nature of the verbal constructions. Therefore, we use a more semantic approach, subdividing predicates into simple verbs and compound verbs (without and with aspectual constructions respectively), and syntactic and phrasal conjuncts (non-idiomatic and idiomatic constructions in the presence of nouns/adjectives with the verbs). Conjunct verbs, due to their semantic implications in Hindi have been detailed separately in subsection 5.2.

- **Simple verbs** : are event mentions constructed by using only the verb, or using the verb and a tense marker (verb + inflectional "*tA*"). Simple verbs denote the habitual nature of the event or action. An example of this construction is:

  6. *rAm* *roz* *shAm* *ko* *ghar* ***jAtA*** ***hai***
     Ram daily evening (acc.) home goes is
     "Ram goes home everyday in the evening"

- **Verbs with auxiliaries**: Verbal auxiliaries include aspect markers and modal indicators that give more information about the event itself. These can be *inchocative predicates* (predicates that de-

note coming to existence of a situation), *aspectual constructions* or simply denoting the possibility of an event. These may also occur in combination, such as *modal aspectual construction*.[5]

7. *SIlA* *kAm* *par* ***jA*** ***rahi*** ***hai***
Sheela work to go cont-fem is
"Sheela is going to work."

5.2 Conjunct Verbs and Phrasal Conjuncts

In Hindi, a conjunct verb is a complex predicate of the construction "Noun/Adjective + Verb" (Begum et al., 2011a). This construction raises the pertinent question of when a noun should be included in the event span while annotating the verbal event. Nouns, when occurring without a postposition with a verb, occur in one of the following contexts: as a participant, as a conjunct verb, and as part of an idiom or phrase.

- **Conjunct Verbs**

  In Hindi, a noun can be "verbalized" by addition of a verbalizer to denote action. The resultant multi-word expression, will behave exactly like a verb, with an argument structure, and temporal and aspectual characteristics of the verb, but the semantic characteristics and meaning of the noun.

  In the example below, *madad* (help), a noun combines with the verb *kI* (to do) to form the event core:

  8. *rAma* *ne* *SyAma* *kI* ***madad*** ***kI***
  Ram (erg.) Shyam (gen.) help did
  "Ram helped Shyam."

- **Phrasal Conjuncts**

  An event mention as a phrasal conjunct includes a noun, and a verb which is *not* a verbalizer. Both the noun and the verb combine to form the event nugget. While the concept being represented by the phrasal event is realized by embedding the noun, the verbalizer presents the concept as an action. In the example below *paisA KAyA* is an idiom in Hindi which, although literally translates to "eating money", is usually interpreted as "scamming someone of their money".

  Unlike conjunct verbs, the semantic information in a phrasal conjunct can not be isolated to a single lexical item, rather, it is a combination of the noun and the verb that provide the meaning of the phrase.

  9. *rAma* *ne* *SyAma* *ka* ***paisA*** ***KAyA***
  Ram (erg.) Shyam (gen.) money ate
  "Ram scammed Shyam"

Contemporary literature analyses similar events as light verb constructions in languages like Hindi and Persian, among others (Vaidya et al., 2016), from a syntactic perspective. While the surface structure is important for the annotation of phrasal events, the difference in nomenclature arises from the fact that phrasal events are not only syntactically unique to their conjunct form (Begum et al., 2011b), but also because the task of event annotation involves extracting information about the semantics of the conjuncts, which is not directly implied or studied in light verb constructions. Therefore, in this framework, they have been referred to as phrasal events rather than light verbs.

[5]Some linguists consider the suffix *tA* as an inflectional perfective aspect marker. However that debate does not concern our annotation of the event span. From the perspective of lexical annotation, verbal auxiliaries are included in the event span.

5.3 Nouns

Nouns categorized as events are either those, which are indicative of an action, or those which can be associated to the aspectual or temporal representation of another event. Nouns classified as events are either those that carry a predicate structure as a result of their nominalization or those whose inherent sense is eventive. For named events, we will be annotating the whole name as the event. This is denoted by the existence of all the above markers (participant or setting) and an attached NP that identifies it as a unique occurrence. A noun which is referred to by an event after it, along with an indication of temporality also indicates that the noun is an event.

10.

[***BUk***	***haDtAl***]	*ke*	*daurAn*	*annA*	*hazAre*	*ji*	***bImAr***	***paD***	***gaye***
Hunger	strike	(gen.)	during	Anna	Hazare	sir	sick	fall	went

During the hunger strike Anna Hazare sir fell sick.

The noun phrase almost distinctly denotes the participation of an entity or the spatio-temporal setting of the occurrence of that event. This can be used to identify the trigger of an event, and hence denote it. A noun is considered to be an event if it occurs in one of the two given contexts:

1. Nouns followed by temporal or time related expressions suggesting that they can be used as (or are) referential in nature for other events.

2. Nouns that are indicative of having lasted a period of time. The duration can be in prolonged or can be immediate.

6 Challenges in Guideline Development

For developing our guidelines, we followed an iterative and incremental approach over three phases. In each phase, a version of the guidelines would be drafted that would then be validated by annotating events in accordance with the current version and then manually investigating the annotations to account for any problems or shortcomings. Shortcomings were majorly identified by calculating inter-annotator agreement and analyzing conflicting annotations. Inter-annotator agreement (strict match) was calculated using the Fleiss Kappa score (McHugh, 2012). Once these flaws were detected, the guidelines would then be accommodated to account for them, creating a new version. The cycle would continue until we reached an exhaustive set of guidelines.

Each version of the guidelines was tested on 50 unique news articles from the Hindi newspaper Dainik Jagran[6]. In each phase, the 50 articles were annotated by 4 annotators, all of whom were native Hindi speakers and had basic training in the field of linguistics. All annotation were done using the BRAT annotation tool Stenetorp et al. (2012a)

6.1 Phase 1

- **Draft**: One of the major challenges we faced in creating the first draft of the guidelines was finalizing the definition of events. As our research is focused on open-domain event detection, our preliminary definition of events was inspired by that of TimeML Pustejovsky et al. (2003). At the end of Phase 1, we had version 0.1 of the guidelines. 50 articles were then annotated as per these guidelines.

- **Challenges**: After manually analyzing the annotations, it was detected that copulative constructions contributed to most of the conflicting annotations. We also observed that annotation of event spans was very inconsistent between annotators. Fleiss' Kappa score observed: 0.59.

[6]https://jagran.com

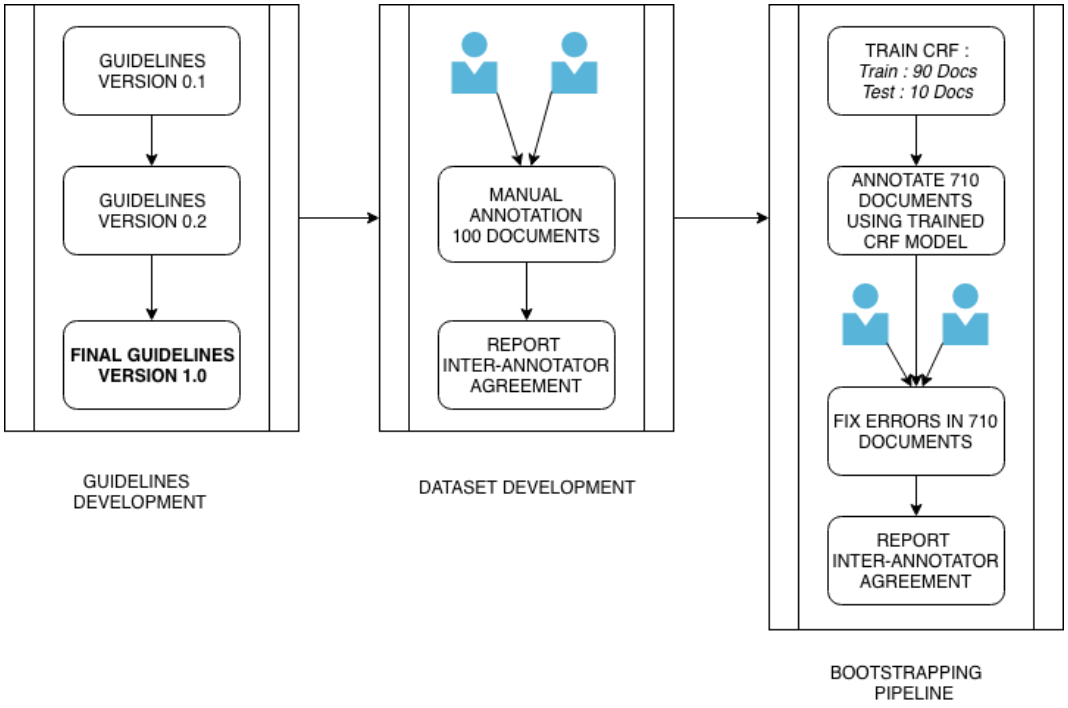

Figure 2: Dataset Development Pipeline

6.2 Phase 2

- **Draft**: During Phase 2, the guidelines were modified to exclude copular constructions from being annotated as events. This was done because copulative constructions do not always highlight a change in the properties of the direct object of the copulative verb and hence. To account for inconsistencies in event span annotation, the guidelines now incorporated rules to mark the tense, aspect and modality (TAM) information of the event within the span of the event (Section 4.2). 50 articles were then annotated as per these guidelines.

- **Challenges**: Manual evaluation of annotations during this phase highlighted three major issues. The first was relation to conflicting annotation of idiomatic expressions (example 1 below). The second being that some annotators had marked the direct object of the event trigger in the event span while the other annotators had not. The final issue showed that Hindi captures certain events (a single semantic idea) in non-contiguous (fragmented) syntactic sequences (example 2 below). Most annotators only marked a part of such constructions. Fleiss' Kappa score observed : 0.72. Examples of conflicting annotations:

11. *rAm se dUr raho, vaha* ***maKan lagAne*** *mein ustAd hai*
Ram from away stay, he/she butter putting in expert is
"Stay away from Ram, he's an expert at flattering people."

12. *unhone* ***ghUs*** *parson subah* ***dI***
They + (erg.) bribe day-before morning gave
"They gave the bribe day before yesterday morning."

6.3 Phase 3

- **Draft**: During Phase 3, the guidelines were further modified to account for the following changes. Idiomatic expressions would be marked events because they derive meaning from ontological roots rather than constructional syntactic uniformity (Butt, 2010). Rules were established to account for annotation of fragmented events (Section 4.3). We now also accounted for events that were and

Data Point	Value
Number of Articles	810
Number of Sentences	13949
Number of Tokens	242201
Number of events	20190
Average Sentence Length (Words)	18

Table 1: Dataset Metrics

Element	Train	Test	Validate
Sent	10318	2430	1201
Tokens	175967	44143	22091
Events	15257	3316	1617
Verbal Events	13808	3061	1480
Nominal Events	1449	255	137

Table 2: Data Split

Features
Word Identity (WI)
Part-of-Speech (POS)
Bi-gram and tri-gram features
Beginning Of Sentence (BOS)
Window features: POS, WI

Table 3: CRF Features

were not syntactically bound with an embedded nominal (Section 5.2). This was the final phase of the guideline development cycle and resulted in Version 1.0 of the Hindi Event Annotation Guidelines. 50 articles were then annotated as per these guidelines.

- **Challenges**: Manual evaluation of annotation during this phase revealed minor discrepancies between annotators. Most of these discrepancies were regarding a missed/skipped annotations. Fleiss' Kappa score observed: 0.86.

7 Dataset

For the task of automated event detection in Hindi, we introduce a gold standard event annotated corpus. This *gold standard dataset*, comprises of 810 event annotated news articles from the financial domain of the Hindi newspaper, Dainik Jagran. The articles have been extracted date from July-December, 2017. The articles in the dataset were manually curated and selected to account for multiple types of events in varying syntactic and semantic conditions. The metrics of the gold standard dataset are described in Table 1. The data was annotated using the BRAT (Stenetorp et al., 2012b) annotation tool.

7.1 Bootstrapping Dataset Development

Once the guidelines were finalized (Version 1.0), four annotators annotated 100 news articles (D1) in accordance with them. We observed a strong inter-annotator agreement, with a Cohen's Kappa score (Cohen, 1960) of 0.84 (strict match) which re-affirmed our confidence in the clarity and coverage of the guidelines.

To further expand our dataset, we use a bootstrapping approach defined in figure 2. Using 90 out of the 100 articles from D1, we train a linear chain CRF to predict events in Hindi text. Features used to train the CRF are shared in table 3. The CRF is then tested on the remaining 10 articles and reports an F1 score of 65.22. We used the trained CRF to annotate the another 710 news articles (D2). Given the low F1 score of the CRF, we were aware that the annotations were erroneous. The same team of annotators that worked on D1, now manually reviewed and resolved all errors in the annotations done by the CRF. The annotators showed a final inter-annotator agreement of 0.79. At the end of this we had 810 articles annotated for events in Hindi (D3), which formed our gold standard dataset. The dataset will be made publicly available on an easily accessible platform upon validation by the community.

8 Conclusion and Future Work

In this paper, the concept of event detection is introduced for Hindi text along with a comprehensive set of annotation guidelines and specifications for detecting events in Hindi text. We introduce an event annotated dataset for Hindi news articles, which is the first dataset in an Indo-Aryan language of this kind to the best of our knowledge. The guidelines presented allowed the annotation of the 810 article dataset with high agreement among annotators, indicating the robustness of the annotation scheme.

This paper is an attempt to preliminarily establish this new direction of NLP research in Hindi. This task of event detection can now be introduced to other Indo-Aryan languages with ease, including Urdu, Bengali, Punjabi, Marathi, Oriya and so on, with minimal changes to the event detection guidelines. These guidelines can therefore be applied for event mention annotation over a family of languages which are low-resource in nature.

The dataset introduced can be used for further tasks such as stative event detection, event classification, and annotating event-entity and event-event relations. This information can be used for tasks like factoid question answering, extractive summarization and other related tasks.

The ISO-TimeML annotation mechanism can be adopted in order to create a TimeBank for Hindi as has been done for the other languages mentioned above, but that is beyond the scope of this paper.

References

Bach, E. (1981). On time, tense, and aspect: An essay in english metaphysics.

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

Begum, R., K. Jindal, A. Jain, S. Husain, and D. M. Sharma (2011a). Identification of conjunct verbs in hindi and its effect on parsing accuracy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 29–40. Springer.

Begum, R., K. Jindal, A. Jain, S. Husain, and D. M. Sharma (2011b). Identification of conjunct verbs in hindi and its effect on parsing accuracy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 29–40. Springer.

Bejan, C. A. and S. Harabagiu (2010). Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1412–1422. Association for Computational Linguistics.

Bharati, A. and R. Sangal (1993). Parsing free word order languages in the paninian framework. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 105–111. Association for Computational Linguistics.

Bittar, A., P. Amsili, P. Denis, and L. Danlos (2011a). 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*, pp. 130–134. Association for Computational Linguistics.

Bittar, A., P. Amsili, P. Denis, and L. Danlos (2011b). 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*, pp. 130–134. Association for Computational Linguistics.

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, 48–78.

Caselli, T., V. B. Lenzi, R. Sprugnoli, E. Pianta, and I. Prodanof (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*, pp. 143–151. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement 20*(1), 37–46.

Doddington, G. R., A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the 4th Language Resources and Evaluation Conference*.

Goyal, K., S. K. Jauhar, H. Li, M. Sachan, S. Srivastava, and E. Hovy (2013). A structured distributional semantic model for event co-reference. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Volume 2, pp. 467–473.

Im, S., H. You, H. Jang, S. Nam, and H. Shin (2009). Ktimeml: specification of temporal and event expressions in korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 115–122. Association for Computational Linguistics.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica 22*(3), 276–282.

Mitamura, T., Y. Yamakawa, S. Holm, Z. Song, A. Bies, S. Kulick, and S. Strassel (2015). Event nugget annotation: Processes and issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 66–76.

Moens, M. and M. Steedman (1988). Temporal ontology and temporal reference. *Computational linguistics 14*(2), 15–28.

O'Gorman, T., K. Wright-Bettner, and M. Palmer (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)*, pp. 47–56.

Pustejovsky, J. (1991). The syntax of event structure. *cognition 41*(1-3), 47–81.

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

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

Song, Z., A. Bies, S. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma (2015). From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 89–98.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii (2012a). 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*, pp. 102–107. Association for Computational Linguistics.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii (2012b). 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*, pp. 102–107. Association for Computational Linguistics.

Vaidya, A., S. Agarwal, and M. Palmer (2016). Linguistic features for hindi light verb construction identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1320–1329.

Vendler, Z. (1957). Verbs and times. *The philosophical review 66*(2), 143–160.

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