

Automatic Identification of Conjunct Verbs in Hindi and Its Application

Ashish Jain (200701012)
Karan Jindal(200701037)

1. Introduction:

a. Problem Statement: This project is a work on identification of conjunct verbs in Hindi. We begin with a set of linguistic diagnostics and see its usefulness in manual identification as well as in building an automatic identification tool for conjunct verbs. We then use the output of this automatic system as a feature in a graph based dependency parser and show an improvement over the state-of-the-art parsing results. We first investigated which noun-verb combination makes a conjunct verb in Hindi. We will then see which of these diagnostics can be used as features in a MaxEnt based automatic identification tool. Finally we will use this tool to incorporate certain features in a graph based dependency parser and show an improvement over previous best Hindi parsing accuracy.

b. Conjunct Verb: There are certain verbs that need other words in the sentence to represent an activity or a state of being. Such verbs along with the other words, required for completion of meaning, are together called Complex Predicates (CP). Complex Predicates exist in great numbers in South Asian languages. Nouns, adjectives and verbs combine with verbs to form complex predicates (CP). The verb in the CP is referred as light verb and the element that the light verb combines to form a CP is referred as host. Butt says that in Hindi/Urdu, the light verb is taken as a contributing SEMANTIC STRUCTURE which determines syntactic information such as case marking whereas host contributes the SEMANTIC SUBSTANCE, i.e. most of the meaning the complex predicate has. Butt has talked about four types of complex predicates: (a) In Syntactic Complex Predicates, the formation takes place in the syntax. (b) In Morphological Complex Predicates, a piece of morphology is used to modify the primary event predication. Well known example is morphological causatives. (c) Light Verbs crosslinguistically do not always form a uniform syntactic category. They are not always associated with a uniform semantics, but they always muck around with the primary event predication. (d) In Semantics, complex predicates represent the decomposition of event structure.

In CPs, Noun/Adjective+Verb combinations are called conjunct verbs and Verb+Verb combinations are called compound verbs. Our work in this project will focus on conjunct verbs in Hindi and their identification.

Examples of N+V combinations: 'वणन + कर', varNan kar, "description + do"

A+V combinations: उपलब्ध है upalabdh hai "available+ be"

2. Previous Work:

There has been quite a few work done so far on the multiword expression identification for Hindi Language. Following are some important and recent work in this area of research:

- **Hindi Compound Verbs and their Automatic Extraction:**

This work analysed Hindi complex predicates and propose linguistic tests for their detection. This analysis enables to identify a category of V+V complex predicates called lexical compound verbs (LCpdVs) which need to be stored in the dictionary. Based on the linguistic analysis, a simple automatic method has been devised for extracting LCpdVs from corpora.

Five different types of V+V sequences in Hindi. These are:

1. V1 stem+V2: maar Daalnaa (kill-put) 'kill'.
2. V1 inf-e+lagna: rone lagna (cry-feel) 'start crying'.
3. V1 inf+paRna: bolnaa paRaa (say-lie) 'say'.
4. V1 inf-e+V2: likhne ko/ke lie kaha 'asked to write'.
5. V1-kar+V2: lekar gayaa 'took and went'

Following diagnostic tests to identify CPs in Hindi:

1. Scope of adverbs
2. Scope of negation
3. Nominalization
4. Passivization
5. Causativization
6. Movement

The tests above have been exhaustively applied to varied data. The results of these tests show that some V+V sequences function as single semantic units and others do not.

- **Relative Compositionality of Noun+Verb Multi-word Expressions in Hindi:**

This work analyses in detail the types of Noun+Verb expressions in Hindi. Then they propose an approach to measure Noun+Verb relative compositionality automatically using maximum entropy model (MaxEnt). MaxEnt integrates various features representing the properties of the Noun+Verb expressions in Hindi. Some of the features used by the MaxEnt are computed by mapping them to Verb-Noun expressions in English.

Following are some of the features which can be used to measure the compositionality of the N+V expressions. The features can be classified into three categories (1) Lexical (word based features like f1, f2, f3), (2) Collocation based (f4, f5, f6) and (3) Contextual (f7). Values of few of the features are calculated indirectly by literally translating N+V expressions in Hindi to Verb-Noun expressions in English.

- **Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora:**

Due to the lack of availability of corpus for identifying the Complex predicates this work highlights the development of first such database based on the simple idea of projecting POS tags across an English-Hindi parallel corpus. The CP types considered include adjective-verb (AV), noun-verb (NV), adverb-verb (Adv-V), and verb-verb (VV) composites. CPs are hypothesized where a verb in English is projected onto a multi-word sequence in Hindi. While this process misses some CPs, those that are detected appear to be more reliable (83% precision, 46% recall).

In this work they first constructed the first corpus-based lexicon of CPs in Hindi based on projecting POS tags across parallel English-Hindi corpora. While such approaches sometimes

leave out some CPs, the ones that are identified are seen to be quite robust. As a result, this appears to be a good first approach for identifying the majority of CPs along with usage data. Moreover, since the language specific input in the procedure is minimal, it can be easily extended to other languages with similar multi word expressions.

3. Rules:

- **Dependency Annotation of Conjunct Verb:**

We have followed Paninian Grammatical framework for dependency annotation of Hindi sentences. We follow the dependency tagging scheme proposed by Rafiya et.al in IJCNLP(2008) for the development of a dependency annotation for Indian Languages.

The noun/adjective and the verb of the conjunct verbs are kept in two separate chunks. The dependency relation of noun/adjective with verb will be pof ("part of" relation), i.e. the noun or an adjective in the conjunct verb sequence will have a POF relation with the verb. We extracted the Noun+Verb combinations which are marked as pof from Hindi treebank and analysed the statistics of the verbs that occur most in conjunct verbs. They are: (1) kar 'do': 958; (2)ho 'be': 274; (3)de 'do':123; (4)le 'give':67.

However the identification of conjunct verbs in Hindi remains an issue of discussion. Since the main problem that is faced in conjunct verb identification is that given a Noun+Verb combination, whether the Noun is part of the CP or is it an overt argument of the verb? Many works in the past have posited a number of diagnostics for identifying conjunct verbs in Hindi.

- **Diagnostics:**

The following are some of the diagnostics mentioned in the literature for deciding which Noun+Verb combinations are conjunct verbs:

I. Coordination Test (D1): This test shows that nouns of conjunct verb don't allow coordination. However it is possible to conjoin the entire N+V combination.

- (3)*log pratiyogita meN rucii aur bhaag le rahe the
 People competition in interest and participation take Prog be-Past
 „People were taking interest and participation in the competition.
- (4) log pratiyogita meN rucii le rahe the aur bhaag le rahe the
 People competition in interest take Prog was and participation take Prog was
 „People were taking interest and participation in the competition.

Example (3) is ungrammatical because rucii 'interest' and bhaag 'participation' are conjoined by aur 'and', whereas these nouns are part of CP. Sentence (4) is grammatical because here the N+V combination i.e., rucii le 'take interest' and bhaag le 'participate' has been conjoined with aur 'and'.

II. Constituent Response Test (Wh-Questions) (D2): CP internal nouns can't be questioned. Only N+V combination can be questioned.

- (5) raam ne jamhaaii lii
 ram Erg yawn take-Past
 „Ram yawned.
- (6)*raam ne kya lii?
 ram Erg what take-Past
 „What did Ram take?

(7) raam ne kya kiya?
raam Erg what do-Past
'What did Ram do?'

Example (6) is ungrammatical because only noun of CP i.e., jamhaai 'yawn' given in example (5) has been questioned. Whereas in (7), the N+V combination, jamhaai le'take yawn' has been questioned.

III. Relativization (D3): CP internal nominals cannot be relativized.

(8)*vah snaan [jo bahut pavitra hai] raam ne gangaa taT par kiya
that bath which lot pure is ram Erg ganga bank on do-Past
'The bath which Ram did on the bank of river Ganga is very pure.'

Sentence (8) is ungrammatical because snaan 'bath' which is noun internal to CP has been relativized by the relative clause.

IV. Adding the accusative case marker (D4): CP internal nominal will not allow the accusative marking.

(9)*raam ne us jamhaai ko liya ...
ram Erg that yawn Acc take-Past ...
'Ram took that yawn.....'

Sentence (9) is ungrammatical because jamhaai 'yawn' which is noun internal to CP has taken an accusative case marker.

V. Adding the Demonstrative Pronoun (D5): CP internal nominal will not take Demonstrative Pronoun.

(10)raam ne yah nirdesh diya
ram Erg. this order give-Past
'Ram gave this order.'

In sentence (10), the demonstrative pronoun yah 'this' is modifying the N+V combination i.e., nirdesh diya 'gave order' and not just the Noun, nirdesh 'order'.

4. Features

Each of noun/adjective-verb pair is represented as a vector of following feature set. The features are categorized into three categories (1) Lexical (word based features like f1, f2, f3), (2) Binary features (f4, f5), (3) Collocation based (f6, f7). These features will help in classifying a noun/adjective-verb pair into literal or conjunct verb class.

a. Verb (f1): Some verbs govern whether an object-verb pair is conjunct verb or not as compared to other verbs. They are more likely to occur as light verbs. Example of such a verb is 'kar' (to do) which accounts for large part of conjunct verb expressions. On the other hand verbs like 'chala' (to walk) occur as literal expression in most cases. Hence, verb will be a good feature for classification task.

b. Object (Noun, Adjective) Lexical (f2): Some objects are more biased towards occurring with a light verb as compared to other objects. These objects have high chances of forming conjunct verb expression with a light verb as compared to other objects.

c. Semantic Category of Object (f3): In some of the theoretical work importance of semantic category of a noun/adjective in identifying conjunct verb has been shown. We incorporated this feature for nouns/adjectives by extracting it from the Hindi WordNet. We referred to the first sense of topmost ontological node of a noun/adjective. Some of the possible semantic categories are ‘Artifact’, ‘Abstraction’, ‘State’, ‘Physical Object’ etc. Total semantic categories are 83; noun/adjective will fall into any of these categories, so this will help in case of unknown nouns/adjectives.

For Example: in the expression ‘*viSvAsaGAwa-karana*’ (meaning ‘to betray’), the Semantic type of ‘*viSvAsaGAwa*’ is “Anti Social”.

d. Post-Position Indicator (f4): is a Boolean feature which will indicate whether a noun/adjective is followed by a post position and then verb i.e. a post-position marker is present between noun/adjective and verb or not. Basic intuition behind this feature is that if a noun/adjective is followed by a post position than it’s a possible candidate of being a part of verb argument structure. Hence, possibly the particular noun/adjective-verb pair doesn’t belong to conjunct verb class, as mentioned in diagnostic number 4 (D4) in section3.

e. Demonstrative Indicator (f5): is a Boolean feature indicating presence of DEM before noun/adjective-verb pair. This diagnostic is explained in section3 as D5.

f. Frequency of Verbs corresponding to particular Object (f6): If a noun/adjective is occurring with few verbs than its high probable that the given noun/adjective-verb pair is a multi-word expression. So the frequency of the number of different verbs occurring with a particular object will be a good indicator for conjunct verbs. For example: a noun ‘*svIkAra*’ (to accept) occurs only with two different verbs –‘*kar*’ (to do) and ‘*hE*’ and noun ‘*kAnUna*’ (law) occurs with five different types of verbs –‘*bawA*’ (to tell), ‘*kar*’, ‘*baxala*’ (to change), ‘*lA*’ (to bring) and ‘*paDa*’ (to study). Therefore, ‘*svIkAra*’ is more likely to form a conjunct verb expression.

g. Verb Argument Indicator (f7): This feature computes the average number of post-position occurring before a unique noun/adjective-verb pair. Reason for exploring this feature is that if an expression has large number of post position occurring before it then its verb’s argument structure is likely to be satisfied because each post-position is preceded by a noun/adjective which may potentially be the argument of the verb. Hence this noun/adjective-verb pair is more probable to form a conjunct verb.

5. Data Set

Following two datasets that are part of Hyderabad Dependency Treebank annotated according to CPG framework are used for doing all the experiments. Here is the description of both the datasets.

1. Dataset-1: Has 4500 manually annotated sentences (200k words approx.). It was released as part of the ICON’10 tools contest on Indian Language Parsing. This dataset was used as a training data, for training the maximum entropy model for the purpose.
2. Dataset-2: Has 1800 sentences. It was released as a part of the ICON’09 tools contest on Indian Language Parsing. This dataset was used as a testing data and the same dataset is used for doing all the parsing related experiments.

Training data has around 3749 unique consecutive noun/adjective-verb pairs out of which 1987 are unique noun/adjective and 350 unique verbs. Semantic category of each object is mined from the Hindi WordNet. The language model consisting of trigrams of words is created for training

data, which is later used for extraction of various features. Testing data has 3613 noun/adjective-verb pairs out of which 998 are conjunct verbs and remaining are literal expressions.

6. Maximum Entropy

The features extracted above are used for binary classification of a noun/adjective-verb expression into conjunct verb and non-conjunct verb using the maximum entropy model. Maximum entropy has already been widely used for a variety of natural language tasks, including language modelling text segmentation part-of-speech tagging and prepositional phrase attachment. The maximum entropy model estimates probabilities based on the principle that the model is consistent with the constraint imposed maintaining uniformity otherwise. The constraints are derived from training process which expresses a relationship between the binary features and the outcome. Some of the features on which training is performed are distinct valued features (f1, f2) while others are real valued feature (f6, f7). These features are mapped to binary features. We used maximum entropy toolkit¹ to conduct our experiments.

7. Experiments

The trained system on the corpus of 4500 sentences is tested on 1800 sentences for measuring its accuracy. The binary classification of noun/adjective-verb test expressions into conjunct verbs and non-conjunct verbs are done. We took different set of features (as mentioned above) for our experiments by trial and error method to come up with the best combination of feature set, which gives the best trained model. The best model (combination of 6 features “f1, f2, f3, f4, f5, f6”) gives us the highest accuracy of around 85.28%. For the baseline for our task we included Verb (f1) and Object (f2) as feature. Table2 gives the overview of useful features which helped in improving the accuracy.

Table2 shows that when the semantic feature (f3) was introduced, it lead to an improvement of around ‘0.75%’, which proves the relevance of this feature. Inclusion of both Boolean features f4 and f5 showed a large jump in accuracy of about ‘3.15%’. Recall that f3 and f4 corresponds to D4 and D5 mentioned above. Addition of feature f6 improved our system by ‘0.54%’ showing dominance of particular objects (as discuss during f6 definition) in conjunct verbs. We have not considered features which will show the steep decrease in accuracy, e.g. feature f7 on addition shows a decrease of ‘7.78%’ with respect to the best accuracy reached so far, and moreover it is even less than the baseline also. The reason for large decrease in accuracy is that it is very difficult to learn the complex verb-argument structure (because of mandatory and optional arguments for a verb), which is the primary motivation for including such feature. We define features (f1+f2+f3+f3+f5) and (f1+f2+f3+f3+f5+f6) as System-1 and System-2 respectively.

Table2. Showing system accuracy with different feature set

Feature set	Accuracy
f1 + f2	(81.59)
f1+f2+f3	(82.34)
f1+f2+f3+f4+f5	(84.74)
f1+f2+f3+f4+f5+f6	(85.28)
f1+f2+f3+f4+f5+f7	(77.44)

8. Application

¹ http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

It had been observed that Dependency framework is the better way to analyze morphological rich free word-order languages (MoRFWO) (such as Czech, Turkish, Hindi, etc). Various data driven and hybrid approaches has been tried but still the current state-of-the-art parsing accuracy hasn't reached to a level which is comparable to English. Complex linguistics phenomenon is considered as a most vital factor for low accuracy of Hindi parsing apart from long distance dependencies, non-projective sentences and less corpus size. In past various morphological, semantic and clause boundary features have been tried to give language specific features in data driven parsing. All these features help in increasing the overall Hindi dependency parsing accuracy, but the gap between labeled and unlabeled accuracy is still large. Others works have pointed that error due to complex predicates are significant in Hindi dependency parsing. Recall that in a conjunct verb it is the noun/adjective-verb complex that forms the predicate thereby controlling the argument structure. This means that unlike a sentence with a normal verb the predicate information in a sentence with conjunct verb is distributed.

In this section, we investigate the effect of using conjunct verb specific features on parser accuracy. MST [36], [37] Parser was used to parse sentence, the MaxEnt based tool described in section 5.3 provides the feature. An improvement of 0.39% in label and 0.28% in label attachment accuracy is achieved.

9. Experiments and Results

We considered the MST+MaxEnt setting mentioned in [38] as Baseline for our experiments. All the parsing related experiments are performed on Dataset-2 as described in section 5.1. Using the output of System-1 and System-2 as described in Section-6, we added conjunct verb feature in each consecutive noun/adjective-verb pair in the dataset. Feature is added in the feature column of CONLL [42] format by giving an extra indicator like 'pof' (for conjunct verb) and 'npof' (for non-conjunct verb), which led to an increase in parsing accuracy using MST. We tried following ways of giving this feature to the parser:-

- a) Giving only "pof" feature to Object.
- b) Giving only "npof" feature to Object.
- c) Giving both "pof" and "npof" feature to object.
- d) Giving "pof" feature to both Object and Verb.
- e) Giving "npof" feature to both Object and Verb.
- f) Giving "pof" and "npof" feature to both Object and Verb.

Total number of noun/adjective-verb pairs is 3613 out of which 962 and 942 are marked as 'pof' and remaining as 'npof' by System-1 and System-2 respectively. Last way (f) gives the best result on 10 cross validation. The parsing result is shown in Table3.

Table3. Average LA (Labeled Attachment), UA (Unlabeled Attachment) and L (Label) accuracies on 12-fold cross validation

	LA (%)	UA (%)	L (%)
Baseline	68.77	85.68	71.90
System 1	69.05	85.68	72.29
System 2	68.52	85.04	71.93

Table4. 2nd and 3rd column represents the number of correctly identified 'pof' and 'npof' labels. Baseline-1 and Baseline-2 gives the number of labels that are correctly identified by the Baseline System group into 'pof' and 'npof' labels in comparison to System-1 and System-2 respectively. These stats are the summation of 12 testing set which are tested during 12-fold cross validation.

	'pof' labels	'npof' labels
Baseline-1	715	1628

System-1	715+ 36	1628+ 21
Baseline-2	713	1630
System-2	713+ 42	1630+ 15

10. Observations:

System-1 shows an increase of 0.39% in label and 0.28% in label attachment accuracy, this increase accounts to the 0.3%, 1.87%, 2.94% and 0.43% increase in labels accuracy of ‘k1’, ‘k2’, ‘pof’, ‘k7p’² respectively. These labels occur in the same environment as ‘pof’, hence the confusion. Both the System-1 and System-2 helps in reducing the ‘npof’ label (like ‘k1’, ‘k2’, ‘k7p’ etc.) confusion for those chunks which are given conjunct verb feature, by correctly identifying 21 and 15 more labels compare to baseline respectively as shown in Table4. Similarly, number of correctly identified conjunct verb labels increase by 36 and 42 in System-1 and System-2 respectively. This increase shows the positive effect of giving label specific feature to noun/adjective-verb pairs. Even if there is an increase in both systems output, the overall accuracy of System-2 is less compare to both System-1 and Baseline results. This decrease is because of indirect wrong learning leading to ambiguity between different labels.

11. Conclusions and Future Work

We have analyzed some of the diagnostics for manual identification of conjunct verb and there relevance in automatic identification. We successfully showed the importance of these diagnostics in statistical techniques by observing the significant increase in overall accuracy of identifying conjunct verbs and there positive effect on parsing accuracy. In future we will try to automate behavioral diagnostics (like D1 and D3) on the availability of large corpus. Although some diagnostics like Constituent Response Test (Wh-Questions) cannot be automated, they can give some theoretical grounding to conjunct verb identification and can complement the statistical tool. We tried to include some context through feature like f7, but they didn’t help. Since, additional context proves helpful in many tasks; we will have to explore this feature. The parsing accuracy showed improvement by incorporating the features given by our tool. Other NLP application tasks such as Machine Translation can also be tried.