Automatic Tripartite Classification of Intransitive Verbs

Nitesh Surtani, Soma Paul
Language Technologies Research Centre
IIIT Hyderabad
Hyderabad, Andhra Pradesh-500032
nitesh.surtaniug08@students.iiit.ac.in, soma@iiit.ac.in

Abstract

In this paper, we introduce a tripartite scheme for the classification of intransitive verbs for Hindi and claim it to be a more suitable model of classification than the classical binary unaccusative/unergative classification. We develop a multi-class SVM classifier based model for automatic classification of intransitive verbs into proposed tripartite classes. We rank the unaccusative diagnostic tests for Hindi based on their authenticity in attesting an intransitive verb under unaccussative class. We show that the use of the ranking score in the feature of the classifier improves the efficiency of the classification model even with a small amount of data. The empirical result illustrates the fact that judicious use of linguistic knowledge builds a better classification model than the one that is purely statistical.

1 Introduction

An automatic classification of verbs that are distinct in terms of their syntactic behavior is a challenging NLP task. Some works have been done for automatic determination of argument structure of verbs (Merlo and Stevenson, 2001) as well as automatic classification of verbs (Lapata and Brew, 1999; Schulte, 2000; Schulte, 2006) following (Levin, 1993) proposal. However, automatic sub-classification of intransitive verbs has not been attempted majorly till now. Sub-classification of intransitive verbs has bearing on various NLP tasks such as machine translation, natural language generation, parsing etc. For example, we take here a case from English-Hindi MT system. English uses nominative subject for all kinds of intransitive verbs whereas Hindi uses ergative case marker *'ne'* on subject when the verb is unergative and in perfect tense whereas unaccusative doesn't as exemplified in (1a) and (1b) respectively.

(1) a. English: Ram ran a lot.

Hindi: *raam-ne* *khub* *dauRaa.*
Ram-erg very much run-3 pft

b. English: The glass broke.

Hindi: *glaas* *TuT-aa.*
Glass break-3 pft

Classifying intransitive verbs of (1a) and (1b) into subclasses can result in producing right case marking on the subject in the target language Hindi. In parsing, identifying the subclass of the intransitive verb helps in predicting the position of the subject in the Phrase structure tree. One effort of subclassification of intransitive verbs is described in Sorace (2000) where intransitive verbs are further automatically classified into unergative and unaccusative following Perlmutter's (1978) proposal of Unaccusativity Hypothesis. This paper follows the proposal of Surtani et al. (2011) where it has been argued that a a tripartite classification better classify Hindi intransitive verbs. This paper develops a multi-class SVM classifier based model for the automatic classification of intransitive verbs in the tripartite classification scheme. We propose in this paper two approaches for developing multi-class classifier: (a) a Language dependent Classifier and (b) a Language Independent Classifier.

26th Pacific Asia Conference on Language,Information and Computation pages 446–455

The paper is organized into the following subsections. In Section 2, we present the related works. Section 3 discusses the issues involved in a bipartite classification of intransitive verbs. Section 4 talks about the Data preparation. In Section 5, we introduce the tripartite classification scheme and gives a mathematical formulation of how it captures the distribution better than the bipartite distribution. Section 6 discusses the ranking and scoring of the syntactic diagnostics proposed by Bhatt (2003). Section 7 presents the SVM-based classification model. Section 8 presents the results of the two classification models which are compared in Section 9. Section 10 concludes the paper and discusses the future directions.

2 Related Works

With Perlmutter's proposal of Unaccusativity Hypothesis, the unergative-unaccusative distinction of intransitive verbs has become cross-linguistically a widely recognized phenomenon and the distinction has been shown to exist in many languages including German, Dutch, Hindi etc. Unergative verbs entail a willed or volitional act while unaccusative verbs entail unwilled or non-volitional act. Various language specific tests have been proposed as diagnostics for the distinction of the verbs of these two classes. Bhatt (2003) proposes various diagnostic tests for Indian languages. We have examined the seven tests that Bhatt (2003) has proposed in his work.

(i) Ergative Subjects: Unergatives sometimes allow ergative subjects with an ergative case marker *'ne'* esp. when paired with the right adverbials and compound verbs (as in (2a)). On the other hand, Unaccusatives do not allow ergative subjects (as in (2b)).

(2) (a.) *raam-ne bahut naach-aa.*
3P.M.Sg-Erg a lot dance-Pfv
'Ram danced a lot.'

(b.) **raam-ne bahut ghabraaya.*
3P.M.Sg-Erg a lot panic-Pfv
'Ram panicked a lot.'

(ii) Cognate objects: These are simply the verbs noun form. Unergatives verbs sometime allow for Cognate objects (as in (3a)) whereas Unaccusatives do not allow for cognate objects.

(3) (a.) *raavan-ne bhayaanaka hasii has-ii.*
3P.M.Sg-Erg horrifying laugh laugh-Pfv
'Ravan laughed a horrifying laugh.'

(iii) Impersonal Passives: The impersonal passive deletes the subject of an intransitive verb and reduces its valency to zero. Unergatives allow for the impersonal passive (as in (4a)) whereas unaccusatives do not.

(4) (a.) *thodii der aur jhool-aa jaaye.*
Some time more swing-Pfv go-Sbjv
'Swing for some more time.'

(iv) Past Participial Relatives: Past participial relatives target the internal/theme argument of the verb, if there is one. The past participial relatives on Unaccusatives have an active syntax taking *'hua'* be-Pfv/ *'gaya'* go-Pfv (as in (5b)) whereas unergatives are ungrammatical with past participial relatives (as in (5a)).

(5) (a.) **kal dauR-aa huaa chhaatra*
yesterday run-Pfv be-Pfv student
'The student who ran yesterday'

(b.) *vahaan bandh-aa huaa ladkaa*
there tie-Pfv be-Pfv boy
'The boy who is tied there'

(v) Inabilitatives: Inabilitatives describe the inability of the agent towards an action which applies to the class of verbs that undergo the transitivity alternation. Unaccusatives enter the inabilitative with active syntax (as in (6b)) whereas Unergatives do not (as shown in (6a)).

(6) (a.) **raam-se ramaa nahii has-ii.*
3P.M.Sg-Instr 3P.F.Sg neg laugh-Pfv.f
'Ram couldn't make Rama laugh.'

(b.) *raam-se ghar nahii banaa.*
3P.M.Sg-Instr house neg build-Pfv
'Ram couldnt build the house.'

(vi) Compound Verb Selection: There seems to be a kind of selection between compound verbs and main verbs. The unaccusative compound

verb *'jaa'* go appears most naturally with unaccusatives while Unergatives tend to take transitive compound verbs like *'le'* -take / *'de'*-give / *'daal'*-did and seem unhappy with *'jaa'* go (as in (7a)).

(7) (a.) *raam-ne pahaaR chaD liyaa.*
3P.M.Sg-Erg mountain climb take-Pfv
'Ram climbed the mountain.'

(vii) Unmarked Subjects for Non-Finite Clauses: Non-Finite clauses in Hindi do not permit overt unmarked subjects (as in (8a)). But inanimate subjects of the Unaccusative verbs can appear without an overt genitive.

(8) (a.) *[raam-ka/*raam tez bhaagna]*
*3P.M.Sg-Gen/*Nom fast run*
zaruurii hai.
necessary is
'It is necessary for Ram to run.'

3 Issues Involved in Binary Classification of Intransitive Verbs

In Hindi as well, syntactic behavior of intransitive verbs, in many cases, depends on which subclass the verb belongs to. However, the neat unergative-unaccusative classification breaks down in Hindi when an intransitive verb takes an animate subject whose volitionality is bleached off by the very semantics of the verb. The absence of a clear-cut distinction due to varied behavior of the verbs of same classes has led to abandoning of this strict two-way classification, as reported for various languages such as German (Sorace, 2000; Kaufmann, 1995), Dutch (Zaenen, 1998), Urdu (Ahmed, 2010) etc. Bhatt also supports the observation that the distinction is not clear-cut for the language. Surtani et al. (2011) argues that a clear-cut two way distinction does not work for Hindi. Let us consider the verb *marnaa* 'die'. The subject of this verb can be an animate volitional entity; however volitionality of the subject is suppressed because one apparently cannot exercise one's own will for 'dying'. The syntactic behavior of such verbs becomes unstable. For example, *marnaa* 'die' behaves like unaacusative verbs as it does not take ergative subject as in:

(9) *kal-ke bhUkamp me bahut log-ne* marA.*
Yesterday-Gen earthquake Loc many people-Erg die- 3 pfv
'Many people died in yesterday's earthquake.'

However, it takes cognate object like other unergative verbs as illustrated in the following example, where *'maut'* is the cognitive object variant of the verb *marnaa* 'die'

(10) *wo kutte ki maut marA.*
'He died like a dog.'

Another case is the verb *girnaa* 'fall'. This verb was originally being classified as an unaccusative verb because the subject of the verb is an undergoer undergoing some kind of change of state. When the subject is inanimate, the unaccusativity feature holds; the verb does not occur with adverb of volitionality as is true for other unaccustive verb. Therefore the following sentence is illegitimate:

(11 a.) **patta jaan-buujh-kar giraa.*
leaf deliberately fall-Pfv
'The leaf deliberately fell.'

However the situation changes when the verb takes an animate human subject. The construction licenses adverb of volitionality as illustrated below:

(11 b.) *raam jaan-buujh-kar giraa.*
ram.M.Sg deliberately fall-Pfv
'Ram deliberately fell.'

With an animate subject, the verb also allows impersonal passive like unergative verbs as shown below:

(11 c.) *calo, eksaath giraa jaaye.*
move, together fall dgo-Sbjv
'Come let us fall down together.'

These verbs taking animate non-volitional subjects [+Ani -Vol] show some properties of unergatives and some properties of unaccusatives. Due to their fuzzy behavior, it becomes hard to classify these verbs. We discuss in Section 5 why it becomes important to keep such verbs in a separate class.

4 Data Preparation

For the preparation of the data for training and testing the model, we have selected a set of 106 intransitive verbs of Hindi and have manually classified them into the proposed tripartite classification

scheme. We have applied seven unaccusativity diagnostics (as discussed in Section 2) on each verb. But due to the polysemous nature of intransitive verbs, the total number of instances rises to 134.

4.1 Polysemous Nature of Intransitive Verbs

While working with intransitive verbs we observe that verbs are highly polysemous in nature. The same verb root might take different kind of subject as a result of which its semantic nuance changes. That affects its syntactic behavior as well. Let us illustrate the case with verb *uR* -'fly'. It can take an animate and also an inanimate subject as shown below:

(12) a. *pancchii uR raha hai.*
The bird is flying.

b. *patang uR gayi.*
The kite is flying.

The difference in animacy of subject determines that the verb in (a) can occur in inabilitative mood while that is not true for the second use of verb as illustrated below:

(13) a. *Pancchii se uRa nahin gaya.*
The bird was unable to fly.

b. **Patang se uRa nahin gaya.*
The kite was unable to fly.

Verb	Gloss	Ergative case?	Cognate object?	Impersonal Passives?	Past Participial?	Inabilitatives with active syntax?	Light verb selection	Overt genitive marker?	Class
uR	fly	Yes	Yes	Yes	No	No	Yes	Yes	1
uR	fly	No	No	No	Yes	Yes	No	No	3

Table 1: Polysemous nature of verb

Since animacy is an important factor for determining subclasses of intransitive verbs we will consider the polysemy of the kind instantiated above as different instances of verbs. Going by that, we have applied the diagnostics on 106 verbs but on a total number of 134 verb instances.

4.2 Training Data

Table 2 presents three instances of our training data. The results of the seven diagnostic tests applied to three intransitive verbs i.e. *jump, sink, get build* , each belonging to a different class of the tripartite scheme are shown.

Verb	English gloss	Ergative case?	Cognate object?	Impersonal Passives?	Past Participial?	Inabilitatives with active syntax?	Light verb selection	Overt genitive marker?	Class
kUDa	jump	Yes	Yes	Yes	No	No	Yes	Yes	1
DUba	sink	No	No	No	Yes	Yes	No	Yes	2
baNa	get build	No	No	No	Yes	Yes	No	No	3

Table 2: Diagnostic applied on Verbs

The corresponding feature values are obtained from the results of these diagnostic tests, maintaining the original unergativity/unaccusativity distinction. The feature corresponding to each diagnostic test is assigned a value 1 in case a instance shows unergative behavior for that diagnostic test and a feature value -1 in case the instance behaves as an unaccusative for that diagnotic test. As already discussed in Section 2, Unergatives take a Ergative case marker, occur in Cognate object, form impersonal passives, do not form part participial, their inabilitatives do not occur with active syntax, select *'le'* -take and *'de'*-give in compound verb formation and take a Overt genitive marker. So, considering the first instance i.e. 'jump' in Table 2, we find that it behaves as unergative for each diagnostic test, and correspondingly is assigned value 1 for each feature. Similarly, the third instance 'get build', behaves as unaccusative for all the diagnostic tests,is assigned value -1 for each feature. The second instance 'sink', belonging to class 2, behaves as unaccusative for first 6 diagnostic tests but as unergative for the last diagnostic test. Table 3 below shows the feature vectors of these 3 intransitive verbs.

Verb	English gloss	Feature Values							Class
kUDa	jump	1	1	1	1	1	1	1	1
DUba	sink	-1	-1	-1	-1	-1	-1	1	2
baNa	get build	-1	-1	-1	-1	-1	-1	-1	3

Table 3: Diagnostic applied on Verbs

The results of the diagnostic tests are used as features for training and testing the SVM model.

5 Triparite Classification

On applying the diagnostics on the intranstive verbs, we make the following observation:

(i) There is no single distinguishing criterion for sub-classifying Hindi intransitive verbs. Some diagnostic tests, however, perform better than the other giving more accurate results.

(ii) Although all the classes tend to show some inconsistency with the diagnostics, verbs taking animate non-volitional [+Ani -Vol] subjects perform most fuzzily. They are therefore most difficult to classify.

The primary purpose of any classification model is to cluster the elements showing similar behavior within same group so that one can predict the properties of the element once its class is known. The aforementioned observations motivate us for introducing a tripartite classification scheme, as a better classification model for classifying the intransitive verbs for Hindi. The major problem in the classification of the intransitive verb is because of the fuzzy behavior of the verbs that take [+Ani -Vol] subejcts. But the number of such verbs is fairly low (15.7%). The unaccusative class approximately covers all the verbs taking non-volitional i.e. [+Ani -Vol] and [-Ani -Vol] subjects and comprises of 67% of the total intransitive verbs. Thus, even after classifying the verb to the unaccusative class, one is not able to predict its properties as one does not know whether that verb belongs to the fuzzy group. The major drawback of the binary classification model is that the complete class of unaccusative suffers because of a fairly small number of fuzzy verbs. On the other hand, a tripartite scheme provides more confidence in predicting the behavior of the intransitive verbs than the binary classification as we are able to predict the behavior of large portion of verbs. We mathematically show that the tripartite model handles the distribution of the intransitive verbs better than the bipartite model for our data.

5.1 Tripartite Classification Scheme

The intransitive verbs are classified in the following manner under the tripartite classification scheme:

Class 1. Verbs that take animate subject and agree with adverb of volitionality.
Property: [+Vol +Ani]

Class 2. Verbs that take volitional animate subject but are not compatible with adverb of volitionality.
Property: [-Vol +Ani]

Class 3. Verbs that take non-volitional subject.
Property: [-Vol -Ani]

5.2 Does Tripartite Distribution Fits Our Data Well?

A distribution model with higher scatter among the classes and low scatter within the classes is considered to be a better distribution, as it ensures that all the classes are well separated and the instances within a class are close to each other. In order to show that the tripartite classification model handles the distribution better than the bipartite model, we use a mathematical formulation which maximizes the inter-class scatter and minimizes the intra-class scatter. The F-test in one-way ANOVA (ANalysis Of VAriance) technique is used for this. It compares the models by determining the scatter within the class and across the class, and the model that maximizes the F-score i.e. the one that has a higher scatter among the class and low scatter within the class, is identified as the model that best fits the population. The feature vectors corresponding to each verb, after the ranking of the diagnostic tests, as shown in Table 6 are used for calculating the F-score.

F-Score: It is the ratio of the measures of two spreads:

$$\mathbf{F} = \frac{MSTr}{MSE} = \frac{Between - sampleVariation}{Within - sampleVariation}$$

MSTr: MSTr (*mean square treatment*) provides a measure of the spread **among** the sample means $\bar{x_1}$, $\bar{x_2}$,.....,$\bar{x_k}$ **(between-sample variation)** by providing a weighted average of the squared differences between the sample means and the grand sample mean $\bar{\bar{x}}$.

$$MSTr = \frac{SSTr}{k-1}$$

where,

$$SSTr = n_1(\bar{x_1}-\bar{\bar{x}})^2+n_2(\bar{x_2}-\bar{\bar{x}})^2+\cdots+n_k(\bar{x_k}-\bar{\bar{x}})^2$$

$$= \sum_{i=1}^{k} n_i(\bar{x_i} - \bar{\bar{x}})^2$$

and n_1,n_2,....,n_k are the *k* samples, $\bar{x_1}$,$\bar{x_2}$,.....,$\bar{x_k}$ are the *k* sample means, and $\bar{\bar{x}}$ is the average of all the $n = n_1 + n_2 + \cdots + n_k$ observations.

MSE: MSE *(mean square error)* provides a measure of the spread **within** the *k* populations **(within-sample variation)** by providing a weighted average of the sample variances ${S_1}^2$,${S_2}^2$,.....,${S_k}^2$ **(within-samples variation)**:

$$MSE = \frac{SSE}{n-k}$$

where,

$$SSE = \sum_{j=1}^{n_1}(x_{1j} - \bar{x_1})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x_2})^2 + \cdots + \sum_{j=1}^{n_k}(x_{kj} - \bar{x_k})^2$$

$$= (n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2 + \cdots + (n_k - 1){S_k}^2$$

$$= \sum_{i=1}^{k}(n_i - 1){S_i}^2$$

where x_{ij} denotes the *j*th value from the *i*th sample.

The results of the F-test are shown below in Table 4.

	BIPARTITE DISTRIBUTION		TRIPARTITE DISTRIBUTION		
	Unaccusative	Unergative	Class1	Class2	Class3
No. of samples	90	44	52	21	61
Sb_i	0.374	0.183	0.415	0.080	0.404
Sw_i	27.407	9.407	10.851	1.208	15.611
SST	0.5569		0.8985		
MSTr	0.5569		0.4492		
SSE	0.8622		0.5333		
MSE	0.0065		0.0041		
F-Score	85.2645		110.3417		

Table 4: Binary Vs Tripartite model statistics

where

$$Sb_i = \sum_{i=1}^{n_k} n_i(\bar{x_i} - \bar{\bar{x}})^2$$

$$Sw_i = \sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2$$

In the tripartite classification model, the within-class scatter for the class is low. Although Class2 has considerable intra-class variability, but the small percentage of verbs in that class doesn't affect the overall MSE value. On the other hand, the unaccusative class has high intra-class variability, and having a high number of intransitive verbs, it affects the MSE value significantly. As shown in Table 4, the higher value of inter-class scatter (MSTr) for the bipartite distribution is compensated by the large MSE value of its distribution. Thus F-score calculated by taking the ratios of MSTr and MSE is higher for the tripartite distribution showing that the tripartite model fits the data better than the bipartite model.

This paper applies a novel computational approach and develops a classification model by employing a Support Vector Machine (SVM) for the automatic classification of intransitive verbs in the triparite classification scheme. We implement two approaches (a) Language Dependent Classifier and (b) Language Independent Classifier for the classification. In order to build the language dependent classifier, i.e., approach (a), we rank the diagnostic tests that Bhatt (2003) has proposed for identifying unergative/unaccusative distinction. These diagnostic tests are in a way checking possibility of occurring of these verbs in various syntactic constructions. We observe that the performance of "Language dependent classifier" is better than the "Language independent classifier" for our data. The classification accomplished by the classifier confirms the fact that verbs of class 2 perform most inconsistently. Thus the paper argues that the model developed in this paper can be used for a better classification of intransitive verbs. The next section proposes a method for ranking the diagnostics proposed by Bhatt (2003).

6 Ranking and Scoring the Diagnostics

We have observed that some diagnostic tests (as discussed in Section 2) are more trustworthy than others in the sense that they can more accurately classify verbs in their respective class. One such test is "Impersonal passive". Most verbs that form impersonal passives are unergative. Such tests are assigned high score which are used as features in developing the classifier model of approach (a). We evaluate the direct correlation of the diagnostic test on the performance of the model in the following manner:

- If the performance of the learning model is largely affected on removal of a diagnostic test

as a feature of the model, then that diagnostic is more important for the model. This entails that introduction of that diagnostic test in the feature vector of the model increases the models accuracy, and hence is more significant for the model.

Table 5 shows the results of the model on pruning the particular diagnostic as feature from the model. The accuracy of the model without the pruning of features is calculated to be 87.42% which is shown to reduce in every case in Table 5. This is because every diagnostic test is adding some useful information to the model. The overt genitive (Unmarked subjects for non-finite clauses) diagnostic seems to perform best for the model on pruning of which the baseline accuracy is reduced by 6.82%.

A diagnostic whose removal from the feature vector affects the model more is regarded to be the better diagnostic test for the model. The %effect on the performance of the model on removal of the diagnostic test is used to calculate the rank and score for the diagnostic. A diagnostic with a better rank is supposed to achieve a higher score. We calculate the score of the diagnostic using the formula:

$$Score = \frac{E_p}{E_b}$$

where
Ep=%Effect on accuracy on pruning the diagnostic
Eb= Accuracy of the model without removal of any diagnostic (87.42%).

Diagnostic	Model Performance On Pruning	%Effect on Accuracy	Rank	Score
Ergative Subjects	84.33	3.09	3	0.03534
Cognate Objects	86.57	0.85	5	0.00972
Impersonal Passives	82.09	5.33	2	0.06097
Past Participial Relatives	86.57	0.85	5	0.00972
Inabilitatives	85.82	1.60	4	0.01830
Compound Verb Selection	85.82	1.60	4	0.01830
Overt Genitive	80.60	6.82	1	0.07801

Table 5: Diagnostic Rank and score

These scores are used to design a new feature vector for the model of approach (a). The feature values corresponding to each diagnostic are multiplied with the corresponding diagnostic scores as shown in Table 6. This feature vector captures the relative significance of the diagnostic with a more relevant diagnostic having a higher ability in unergative/unaccusative distinction. The comparison between the two models: the one which incorporates the ranking score information and the other which do not has been discussed in Section 11.

Verb	Gloss	Feature Values							Class
kUDa	jump	0.035	0.01	0.061	0.01	0.018	0.018	0.078	1
DUba	sink	-0.035	-0.01	-0.061	-0.01	-0.018	-0.018	0.078	2
baNa	get build	-0.035	-0.01	-0.061	-0.01	-0.018	-0.018	-0.078	3

Table 6: Feature vector after Ranking

7 Classification Model

We employ a multiclass SVM as a computational model to analyse behavior of the intransitive verbs. In response to the observations, we use the model to show that Class2 samples are indeed hard to classify with maximum misclassification rate. On the other hand, Class1 and Class3 verbs are classified quite well with a low misclassification rate. We develop two models: (a) Language dependent Classifier which takes a feature vector that incorporates diagnostic scores (as shown in Table 6) and (b) Language Independent Classifier which takes feature vector with binary values as shown in Table 3. We then compare the two models, one without prior diagnostic rank information and the other incorporating the relative linguistic significance of the diagnostic by calculating the ranked diagnostic scores, and show that the ranked model outperforms the one without ranking information. This learned model can also be applied for the classification of new intransitive verbs.

7.1 Support Vector Machine

Support vector machines, (Vapnik, 1995), are computational models used for the classification task in a supervised learning framework. They are poular because of their good generalization ability, since they choose the optimal hyperplane i.e. the one with the maximum margin and reduces the structural error rather than empirical error. Kernel-SVMs, (Joachims, 1999), are much more powerful nonlinear classifiers which obtain a maximum-margin

hyperplane in a transformed high (or infinite) dimensional feature space, non-linearly mapped to the input feature space. Although SVMs are originally designed for binary classification tasks, they are extended for building multi-class classification models. We use LIBSVM library (Chang and Lin, 2011) which implements the "one-against-one" approach for multi-class classication. The next section describes the implementation of the SVM model for classifying intransitive verbs of Hindi.

7.2 Pre-processing

Before performing the experiments, first the data is preprocessed by centering the mean for each feature. Mean centered data have a mean expression of zero, which is accomplished by subtracting the feature mean from each data entity.

$$X_M = X - \bar{X}$$

7.3 Training and Testing

Since the data is scarce, so for the better prediction of the error, we use the k-fold cross-validation. The data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently *k* iterations of training are performed such that in each iteration, a different fold of the data is held-out for testing while the remaining *k*-1 folds are used for training. When *k* is equal to the number of samples, there is only one test sample in each experiment and the technique is referred to as *Leave-One-Out* (LOO). The advantage of k-fold cross-validation is that all the samples in the dataset are eventually used for both training and testing. So, the true-error, i.e. that error over the test data is estimated as average error rate.

$$E = \frac{1}{k} \sum_{i=1}^{k} E_i$$

We calculate the average true error, E, for different values of *k*. Table 7 shows the accuracy of the model for different values of *k*. Other model parameters are varied keeping the k constant. We find that the average accuracy of the model is maximum for $k = 15$, for which the true error, E is minimum. Optimal values of other parameters are discussed in Section 7.4. So, $k = 15$ is chosen as the best k value and is used in further calculations.

K	3	5	7	9	12	15	134
Accuracy	77.57	83.06	85.99	86.90	87.19	87.42	86.56

Table 7: Accuracy on different folds

For calculating the class accuracies, the set of 134 intransitive verbs is partitioned into $k = 15$ folds, and each fold is used once for testing while other folds are used for training the model. While testing the model in each iteration, the correctly classified and the misclassified samples of each class are identified. For this experiment, we have taken the optimal model parameters i.e. $C = 1$ and sigmoid kernel function, as discussed in Section 7.4. The numbers of misclassified and correctly classified samples for each class are presented below in Table 8.

7.4 Model Parameters

Two model design parameters i.e. C value and the kernel functions and their ranges after optimization are discussed below.

C Value: C value decides the weight for the rate of misclassification. The accuracy of the classifier at lower value of C is low but it increases drastically on increasing C upto a point and then drops down again on further increment. The value of C has been varied from 0.001 to 100000. It has been represented in log scale in Figure 1.

Kernel Function: Four kernel functions, namely, Linear, Polynomial, Radial Basis and Sigmoid are used in order to tune the model to the best performance for different C values, as shown in Figure 1. The results show that the Sigmoid kernel outperforms other kernel functions.

The optimal parameters of the model are achieved when the value C is set to 1 and kernel function used is sigmoid with a accuracy of 87.42%.

8 Results

Table 8 below represents the number of verbs of corresponding class classified into Class1, Class2 and Class3. So, the diagonal elements of the matrix represent the correctly classified samples and the rest of the samples are misclassified. Correspondingly, the class accuracies are calculated as the ratio of correctly classified verbs and the frequency of that class. The results confirm our motivation that Class2

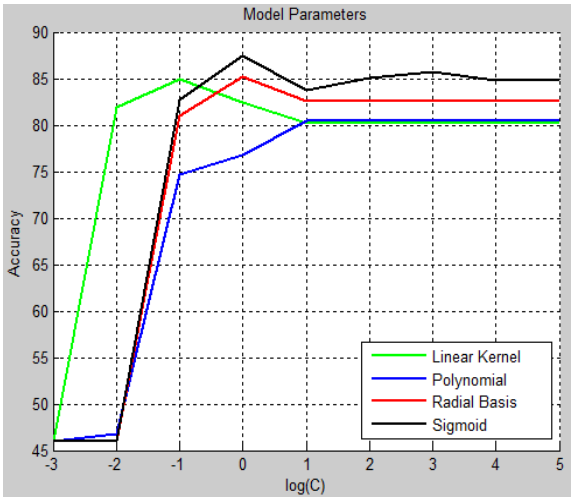

Figure 1: Parameters for Language Independent Model

	Class1	Class2	Class3	Total	Class Accuracy (in %)
Class1	48	4	0	52	92.3
Class2	3	14	4	21	66.67
Class3	1	8	52	61	85.24

Table 8: Results of the model

verbs perform most fuzzily with the highest misclassification rate. This fuzzy behavior of these verbs causes both the unergative and unaccusative classes suffer having low class accuracy in a bipartite classification. A tripartite approach for the classification of intransitive verbs handles this problem efficiently. In the tripartite classification scheme described in this paper, verbs that take [+Ani +Vol] and [-Ani -Vol] subjects are classified in Class1 and Class3 respectively with high class accuracies of 92.3% and 85.24%. The verbs taking [+Ani -Vol] subjects are handled separately in Class2, which has a class accuracy of only 66.67%. The verbs such as *ruk* 'stop', *bhool* 'forget' and *sarak* 'creep' which belong to Class1 are misclassified into Class2 whereas the verbs such as *bacch* 'saved' and *darr* 'scared' are misclassified from Class2 to Class1.

9 Comparison Of The Models

The two classifiers models, (a) Language dependent Classifier and (b) Language Independent Classifier are compared for their performance on the Hindi data. The accuracies of the two models at different values of k are shown in Figure 2. The findings show that the model constructed by approach (a), incorporating linguistic information in terms of relative diagnostic scores, outperforms the model designed using approach (b), the one that doesn't uses any prior linguistic information. Even for smaller values of k, the Language-Dependent model gives a considerably high accuracy showing that the model has good generalization ability and is able to learn a classifier that performs quite well even on small training data. As the number of folds increase, both models attain approximately equal accuracies. The Language Dependent model achieves a maximum accuracy of 87.69% for k=7 when C=100 and kernel function is sigmoid function whereas the Language-Independent model achieves a maximum accuracy of 87.42% for k=15 when C=1 and kernel function is sigmoid function.

Figure 2: Performance of the two models

10 Conclusion

This paper presents a tripartite approach for the classification of intransitive verbs and the results reveal that it does handle the distribution of intransitive verbs better than the binary distribution. The intransitive verbs that take [+Ani -Vol] subjects are most incompatible with the Unaccusativity diagnostics, which are kept in Class2 in our classification scheme. The verbs of this class are most incompatible with the unaccusativity diagnostics and show fuzzy behavior causing a major problem in the unergativity/unaccusativity distinction. With this

observation, we keep these verbs in a separate class so that the other two classes, which perform well over the diagnostic tests, are well separated. The results given by the model reveals that this observation is correct and Class2 verbs indeed show fuzzy behavior with a high misclassification rate. The other two classes have low misclassification rate and have shown to perform quite well. The ranking of the diagnostics with their corresponding scores gives the relative significance of the diagnostic for the unaccusative-unergative distinction of the intransitive verbs. The model incorporating this relative rank information in the form of diagnostic score has shown to outperform the model without that information. The training of the model will be improved by increasing the number of verbs used for training.

As part of the future work, we will explore the applications of the work in Machine Translation systems and Natural Language Generation.

References

Annie Zaenen 1998. *Unaccusatives in Dutch and the Syntax-Semantics Interface*. CSLI Report 123. Center for the Study of Language and Information, Stanford, CA.

Antonella Sorace 2000. *Gradients in auxiliary selection with intransitive verbs*. Language 76, 859890.

Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press, Chicago.

Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Semantics Interface*. Cambridge, MA: MIT Press.

Cengiz Acartrk and Deniz Zeyrek. 2010. *Unaccusative/Unergative Distinction in Turkish: A Connectionist Approach*. the 23rd International Conference on Computational Linguistics. Proceedings of the 8th Workshop on Asian Language Resources. Beijing, China, 2010. pp. 111-119.

Chih-Chung Chang and Chih-Jen Lin. 2011. *LIBSVM : a library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.

David M. Perlmutter. 1978. *Impersonal passives and the unaccusative hypothesis*. Proceedings of the 4th Berkeley Linguistics Society, 157-189.

Deniz Zeyrek. 2004. *The role of lexical semantics in unaccusative-unergative distinction in Turkish*. In Comrie, B. Soloveyv, V., Suihkonen, P (Eds), international Symposium on the Typology of Argument Structure and Grammatical Relations in Languages Spoken in Europe and North and Central Asia LENCA-2). pp 134-135. Kazan State University, Tatarstan Republic, Russia, 2004.

Ingrid Kaufmann 1995. *O- and D-Predicates: A Semantic Approach to the Unaccusative-Unergative Distinction*. Journal of Semantics 12, 377-427.

Maria Lapata and Chris Brew. 1999. *Using Subcategorization to Resolve Verb Class Ambiguity*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 397-404. College Park, MD.

Nitesh Surtani, Khushboo Jha and Soma Paul. 2011. *Issues with the Unergative/Unaccusative Classification of the. Intransitive Verbs*. International Conference on Asian Language Processing (IALP), Penang, Malaysia.

Paola Merlo and Suzanne Stevenson. 2001. *Automatic verb classification based on statistical distributions of argument structure*. Computational Linguistics, 27(3):373408.

Rajesh Bhatt. 2003. *Causativization, Topics in the Syntax of the Modern Indo-Aryan Languages*. Handout.

Sabine Schulte im Walde. 2000. *Clustering verbs semantically according to their alternation behaviour*. In Proceedings of COLING, pages 747753, Saarbrucken, Germany.

Sabine Schulte im Walde. 2006. *Experiments on the automatic induction of german semantic verb classes*. Computational Linguistics, 32(2):159194.

Tasveer Ahmed 2010. *The Unaccusativity/Unergativity Distinction in Urdu*. Journal of South Asian Linguistics, North America.

Vladimir N. Vapnik 1995. *The Nature of Statistical Learning Theory*. Springer.

Thorsten Joachims 1999. *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*. B. Schlkopf and C. Burges and A. Smola (ed.). MIT Press.