Supervised Categorization for Habitual versus Episodic Sentences

Thomas A. Mathew
Department of Linguistics
Georgetown University
Washington, DC 20057, USA
tam52@georgetown.edu

E. Graham Katz
Department of Linguistics
Georgetown University
Washington, DC 20057, USA
egk7@georgetown.edu

Abstract

We implement and evaluate systems for automatically distinguishing habitual and episodic sentences. Using features such as tense, aspect, noun phrase characteristics, temporal modifiers, specific adverb modifiers, and specific verb auxiliaries on genericity we built and evaluated a supervised machine learning classifier that provides 86.3% precision in disambiguating habitual and episodic sentences. This compares against a majority class baseline of 73.1% precision. In order to support these objectives, a representative corpus sample was hand-annotated.

1 Introduction

The distinction between habitual and episodic sentences is well known in the linguistic semantics literature (Dahl 1985, Krifka et al. 1995, Carlson 2005). Habitual sentences, such as (1a), are used to state general facts, while episodic sentences, such as (1b) are used to refer to specific events.

(1) a. *Bombs explode when ignited.*
b. *The bomb exploded.*

Making this semantic distinction is essential for a number of natural language processing tasks, from document summarization to event extraction, but it is not easy. There are few overt features that unambiguously signal habitual/episodic distinction in languages of the world (Dahl, 1995). Fortunately there are many features which provide good evidence for this classification: The English simple present tense, for example, is used primarily with habitual sentences, while definite temporal adverbials such as *last week* are used almost exclusively with episodic sentences.

In this paper we report on our experiments in building a system for automatically classifying sentences as habitual or episodic. This involved the development of a corpus of annotated example sentences, the selection of features for use in classification, and the application of a number of classification algorithms.

1.1 Habitual vs. Episodic Sentences

Habitual sentences come in a variety of sub-types (Rimell 2004): Those with generic antecedents that refer to generalizations about classes of individuals (2ab), and those that refer characteristics of a single individual (2c).

(2) a. *Bears usually eat blackberries*
b. *Italians drink wine*
c. *Jane wakes up at 7:00 AM*

In what follows we take habitual sentences to be sentences whose main verb is lexically dynamic, but which do not refer to particular events. Although stative sentences such as *Italians like wine*, share many of the semantic features of habituals (cf. Chierchia 1995), we exclude statives from our investigation because they do not exhibit the habitual/episodic ambiguity. Stative verbs such as *like* cannot, in general, have an episodic interpretation.

Episodic sentences refer to finite irregular events. Examples are shown in (3) below (Carlson 2005).

(3) a. *Mary ate oatmeal for breakfast this morning*
b. *Each student handed in a completed assignment*

As shown by (3b), episodic sentences can also refer to a plurality of events. The contrast between episodic and generic sentences, then, concerns not the number of events referred to, but rather the gnomic, characterizing or other “intensional” properties of the sentence (Carlson 2005).

Some markers indicate quite clearly that a sentence is habitual. These are frequency adverbs (such as *usually*, *typically*), use of a quantificational temporal modifier (such as *every night*) and use of a habitual past modifier (such as *used to* or *would*). Other features are less direct, such as the forms of the grammatical argument NPs: In (4ab) we see that number of the grammatical object NP can influence the classification, but in (4c) also that this depends on the semantics of the NP itself.

(4) a. *John smoked cigarettes*
b. *Mary smoked a cigarette*
c. *Bill smoked a pipe*

Very few markers indicate clearly that a sentence is episodic. One explicit indicator of episodicity is definite temporal modifiers as shown by the modifier *twice this week* in the example (5) below (Cowper 2003).

(5) *Angus Young wore a school uniform twice this week*

In addition, the progressive in English is typically but not absolutely predisposed towards an episodic interpretation (Krifka et al. 1995). Again, there are important lexical semantic factors that can play a crucial role, as the contrast between examples (6a) and (6b) show (the former being episodic and the latter being habitual).

(6) a. *John is eating an apple*
b. *John is seeing a girl from Brooklyn*

The class of a sentence is not always solely determined by sentence-internal factors; in such cases it is only in the context of a discourse that the sentence can be classified, as in (7), where (7a) is habitual and (7b) is episodic.

(7) a. *John rarely ate fruit. <u>He just ate oranges</u>*
b. *John didn’t eat much at breakfast. <u>He just ate oranges</u>*

In this paper we will be concerned with the problem of determining automatically on the basis of sentence internal information whether a sentence is habitual or episodic. We are leaving cases such as (7) to future work.

1.2 Related Work

While some work has been done on the related tasks of distinguishing generic from specific NP reference (Suh 2006), and of determining the lexical class of a verb (Brent 1990, Siegel 1999) we know of no other study on the supervised classification of habitual/episodic sentences. As with Suh’s task, our task is concerned with classifying particular uses of verbs in context. In that sense we are engaged in a kind of word sense disambiguation undertaking. This distinguishes our work from that of Brent and Siegel, which were concerned with classifying particular lexical verbs. As noted, however, the event/state distinction shares much in common with habituality/episodicity distinction. A number of the features that Siegel used in distinguishing stative from eventive verbs have proven useful in our task. These are listed in Table 2.

2 Annotation and Feature Analysis

In order to investigate the effects of sentence internal features on the classification, we manually classified a number of sentences as to their habituality. We randomly selected sentences from the Penn Treebank (from the WSJ and Brown corpus) for manual annotation; the only restriction applied was that the verb predicate was not lexically stative. We then expanded our data set to include all sentences in the corpus whose verb predicate was a morphological variant of one out of the set of verb forms in the initial sentence group; this was done so as to include the full range of feature and category variation for each verb form. Each of these sentences were reviewed in context and annotated as either episodic or habitual based on the following criteria: 1) We verified if the sentence included features that provided explicit category attribution (see 1.1), 2) We tested whether sentence meaning changed by modifying the verb predicate with *usually* (no change in meaning indicated habituality), and 3) We considered intuitive judgments of the sentence category of the prior and following sentence in the discourse. For all cases where the criteria did not provide attribution, we applied intuitive judgment.

The annotated data comprised 1,816 examples with 72 distinct lexical verbs. As indicated in Table 1, although the vast majority of the examples were episodic, nearly 20% of the cases were habitual. The verbs themselves showed clear biases, with a number of lexical verbs appearing practically only in either one class. In order to minimize the impact of the lexical verb on the classification task and to focus on the grammatical features, we chose to discard verbs that showed an overwhelming bias for either category. Out of the 72 verb base forms, 53 showed greater selection for episodicity - we discarded the top 25% (=14). Likewise, 12 verb base forms showed greater selection for habituality and the top 25% (=3) were discarded. The elimination step reduced the data to 1,052 examples covering 57 verb stems.

	Distribution (%)	
Category	**Before verb processing**	**After verb processing**
Habitual	19.9	26.9
Episodic	80.1	73.1

Table 1. Category Distribution

Feature	**Domain of Values**
Tense	Present, Past, Infinitive
Progressive Aspect	Presence, Absence
Perfect Aspect	Presence, Absence
Quantificational Temporal	Presence, Absence
Specific Temporal	Presence, Absence
Bare-plural Subject	Presence, Absence
Definite Subject	Presence, Absence
Absent Object	True, False
Bare-plural Object	Presence, Absence
Definite Object	Presence, Absence
Conditional	Presence, Absence
'at' Prepositional Phrase	Presence, Absence
'in' Prepositional Phrase	Presence, Absence
'on' Prepositional Phrase	Presence, Absence

Table 1. Features

We then selected syntactic/grammatical markers that were shown to be statistically significantly related to the sentence category. We chose to focus on features that could be derived directly from the Penn Treebank annotation scheme (cf. Rohde 2000). These features are listed in Table 2.

These features were analyzed as to their selectional significance by evaluating the distribution of the category over a particular feature value and comparing the deviation against the baseline distribution (shown in Table 1) to determine whether the feature significantly selects for a particular category or not. To evaluate significance we applied a binomial distribution model. The standard deviation of a particular category in a multinomial distribution is evaluated using the expression below where i is the sentence category (i.e. habitual or episodic), P_i is the probability of the category and n is the total number of sentences.

$$(8) \quad \sigma_i = \sqrt{n\, P_i\,(1 - P_i)}$$

For each category, we determined upper and lower thresholds as being 3 standard deviations from the baseline distribution. Any features which had a category distribution outside the upper limit were considered to show significant positive selection for a category; likewise any features with a category distribution below the lower limit were considered to show positive selection for the opposing category. The upper and lower thresholds were calculated using the expression below where n_i is the number of sentences of category i.

$$(9) \quad \frac{n_i \pm 3\,\sigma_i}{n}$$

Category	n_i	σ_i	**Lower Limit**	**Upper Limit**
Habitual	284	14.4	22.9%	31.1%
Episodic	768		68.9%	77.1%

Table 2. Distribution Outlier Thresholds

The variation of the distribution of sentence category with the different features[1] is shown below. We have used the notation ↑ to indicate that there is likely to be a positive correlation between the feature and the category, ↓ indicates the likelihood of a negative correlation and → indicates the likelihood of there being no correlation.

[1] The detailed feature extraction mechanisms applied are described in Mathew (2009).

Feature		Category	
		Hab.	Ep.
Present Tense	Count	149	29
	% to Feature	↑ 84%	↓ 16%
	% to Category	52%	4%
Past Tense	Count	121	702
	% to Feature	↓ 15%	↑ 85%
	% to Category	43%	91%
Infinitive	Count	14	37
	% to Feature	→ 27%	→ 73%
	% to Category	5%	5%
Progressive Aspect	Count	3	12
	% to Feature	↓ 20%	↑ 80%
	% to Category	1%	2%
Perfect Aspect	Count	11	13
	% to Feature	↑ 46%	↓ 54%
	% to Category	5%	1%
Specific Temporal	Count	9	90
	% to Feature	↓ 9%	↑ 91%
	% to Category	3%	12%
Quantificational Temporal	Count	53	5
	% to Feature	↑ 91%	↓ 9%
	% to Category	19%	1%
Bare-plural Subject	Count	20	9
	% to Feature	↑ 69%	↓ 31%
	% to Category	7%	1%
Definite Subject	Count	200	692
	% to Feature	↓ 22%	↑ 78%
	% to Category	70%	90%
Indefinite Subject	Count	84	76
	% to Feature	↑ 53%	↓ 48%
	% to Category	30%	10%
Absent Object	Count	93	232
	% to Feature	→ 29%	→ 71%
	% to Category	33%	30%
Bare-plural Object	Count	16	18
	% to Feature	↑ 47%	↓ 53%
	% to Category	7%	2%
Definite Object	Count	57	319
	% to Feature	↓ 15%	↑ 85%
	% to Category	20%	42%
Indefinite Object	Count	227	449
	% to Feature	↑ 34%	↓ 66%
	% to Category	80%	58%
Conditional	Count	10	37
	% to Feature	↓ 21%	↑ 79%
	% to Category	4%	5%
'at' Prepositional Phrase	Count	4	26
	% to Feature	↓ 13%	↑ 87%
	% to Category	1%	3%
'in' Prepositional Phrase	Count	15	36
	% to Feature	→ 29%	→ 71%
	% to Category	5%	5%
'on' Prepositional Phrase	Count	4	22
	% to Feature	↓ 15%	↑ 85%
	% to Category	1%	3%

Table 3. Analysis of Features on Sentence Category

The metric '% to Feature' represents the distribution by sentence category for all sentences for which the feature is set. The metric '% to Category' represents how prevalent the feature is amongst the entire set of sentences in that category. The former impacts the precision and the latter impacts the recall of a classifier.

We see that present tense (habitual 84% of the time) and presence of a quantifying temporal (habitual 91% of the time) are the two best indicators for habituality. 80% of all habitual sentences have an indefinite object – however by itself, the presence of an indefinite object is not sufficiently selective for habituality (habitual 34% of the time). It is interesting to observe that single features that explicitly license habituality are not as frequently used to express habituality as compared with constructions which on face value are ambiguous. Past tense and presence of a definite subject provide a good indication on episodicity and are also largely prevalent as a feature among all the episodic sentences identified. Other features such as the presence of a specific temporal adverbial and use of an '*at*'-headed prepositional phrase in either a locative, temporal or directional manner provide good indication of episodicity but are not very prevalent among the set of episodic sentences.

From the analysis we have done, it can be observed that tense as a single feature not only provides the highest overall precision in category discrimination but also that the past and present domain of the tense feature covers a significant population of episodic and habitual sentences respectively, which positively improves recall.

3 Classification Task

In this paper we consider two rule-based classifiers namely Association Rule Mining and Decision Trees, as well as a Naïve Bayes probabilistic classifier as methods to predict the sentence category. We measured classification effectiveness in terms of the classic information retrieval notions of precision and recall for each class. In the context of this study, precision can be thought of as the probability that if a random sentence is classified with a certain category, the decision made is correct. By contrast, recall can be thought of as the probability that if a sentence should be classified with a particular category, the decision is taken.

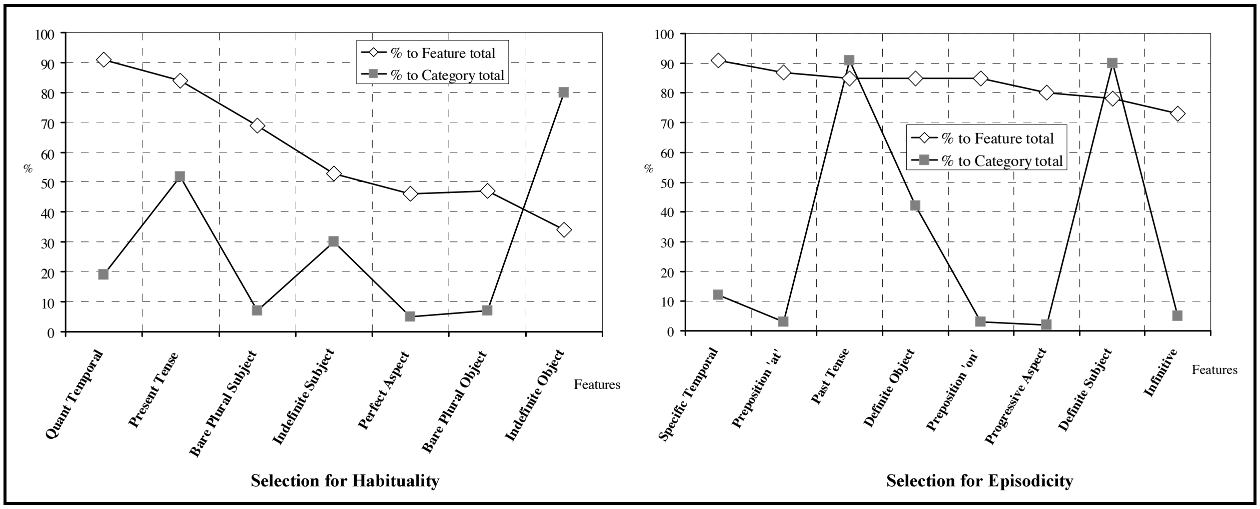

Figure 1. Features selecting for Habituality and Episodicity

The precision and recall performance of our classifiers were compared with a majority-class baseline of 73.1% overall precision with 100% episodic recall and 0% habitual recall.

3.1 Association Rule Based Classifier

Rule based classifiers, which are defined by a disjunctive normal form formula (i.e. a disjunction of conjunctions), have been widely applied in real world applications because of the easy interpretability of rules. Such a classifier is generally created by: 1) Rule discovery by applying multivariate analysis for attribute association, and 2) Pruning to weed out inconsequential rules.

For the step of rule discovery, we applied an association rule mining algorithm namely the Predictive Apriori algorithm (Scheffer 2004). The algorithm proposes a predetermined number association rules which are sorted based on the probability that a rule R selects for a category given that R selects for a feature set. We ran the algorithm to generate 100 association rules which we manually pruned using the criteria: 1) We only included rules which selected for episodicity > 85% of the time or habituality > 80% of the time, 2) We disregarded all rules which identified 5 or less sentences in our data, 3) If rule $R_1 \subset R_2$, we included rule R_1 and discarded rule R_2, and 4) By category, we sorted the rules in descending order of recall and iterated through each of the rules; if the rule set $\{R_1, R_2 .. R_n\}$ and the rule set $\{R_1, R_2 .. R_n, R_{n+1}\}$ covered exactly the same number of sentences for the category they predominantly select for, we considered there to be no information gain provided by rule R_{n+1} and discarded it

After pruning, association rules that indicate habituality more than 80% of the time are shown in Table 5 below. These 4 patterns cover a total of 213 sentences out of which 173 are habitual which gives a habitual precision of 81.2% and a habitual recall of 60.9%. The low recall may be because many indicators of habituality are outside our identified feature set (e.g. discourse or semantic features) or that there exist rare syntactic patterns which select for habituality but which get pruned out because of the sparseness of such patterns in the selected data set. The pruning drops some interesting patterns – for example all examples (=7) where present tense sentences where conditionals are present show 100% selection for habituality.

The pruned set of association rules that indicate episodicity more than 85% of the time are shown in Table 6. These 12 patterns cover a total of 882 sentences out of which 735 are episodic which gives an episodic classification precision of 83.3% and an episodic recall of 95.7%. The pruning again drops some relevant patterns – for example the selectivity for the presence of a specific temporal is ignored.

The high degree of coverage provided by the 16 patterns identified (4 habitual patterns + 12 episodic patterns) provides validation that the features selected are relevant for this categorization task.

Tense	Definite Subject	Definite Object	Specific Temporal	In Prep	At Prep	On Prep	Quant Temporal	Habitual Count	Habituality %	Cumulative Habitual Count
PR	-	-	-	-	-	F	-	148	84	148
PR	-	-	-	-	F	-	-	147	84	149
-	-	-	-	-	-	-	T	53	91	171
-	F	F	F	T	-	-	-	6	100	173

Table 4. Association Rules for Habituality

F	False
PA	Past Tense
PR	Present Tense
T	True
-	Irrelevant

Progressive	Tense	Definite Subject	Bare Plural Subject	Definite Object	Absent Object	At Prep	In Prep	On Prep	Quant Temporal	Specific Temporal	Conditional	Episodic Count	Episodicity %	Cumulative Episodic Count
-	PA	-	-	T	-	F	-	F	-	-	-	269	90	269
-	PA	T	-	T	-	-	-	-	-	-	-	268	91	289
-	-	T	-	T	-	F	-	F	-	-	F	255	86	308
F	-	T	-	T	-	-	-	-	-	F	-	250	85	312
-	PA	-	-	-	T	-	-	-	F	-	-	211	86	523
-	PA	-	F	-	F	-	-	-	-	-	-	209	85	723
-	-	-	-	-	F	-	T	-	F	-	-	29	85	724
-	-	-	-	-	F	-	-	-	F	-	T	28	88	726
-	-	-	-	-	-	T	-	-	-	-	-	26	87	727
-	IN	-	-	-	F	-	-	-	F	-	-	23	85	733
-	-	-	-	-	-	-	-	T	F	-	-	22	92	735

Table 5. Association Rules for Episodicity

Feature	Habitual		Episodic	
	Precision	Recall	Precision	Recall
Association Rule[2]	81.2%	61.9%	83.3%	95.7%
J48	84.3%	60.6%	86.8%	95.8%
Naïve Bayes	81.7%	62.7%	87.3%	94.8%

Table 6. Performance of Classifiers

3.2 Decision Tree Classifier

Mitchell (1997) notes that Decision Tree learning algorithms are appropriate for classification problems that have: 1) attribute-value pair features, 2) a discrete target class, 3) situations where the training data may be error prone, and 4) situations where attribute values are missing. This description fits the data that we are attempting to classify. Similar to the association rule classifier described earlier, decision trees also allow the inner model to be scrutinized. We chose the J48 algorithm (Weka's implementation of C4.5) for our decision tree. For smoothing of the results, we applied ten-fold cross validation while evaluating performance.

The decision tree created by Weka is shown in Figure 2. The two most significant features applied in the decision tree are the presence of a quantitative temporal and the tense of the sentence which are the two highest branching nodes. Altogether there are 13 possible paths to navigate the decision tree out of which the 4 underlined patterns have the largest data coverage and hence largely contribute to the performance of the classifier.

```
quant_temporal = T: HABITUAL (58.0/5.0)
quant_temporal = F
|  tense = PA
|  |  bare_pl_subject = T
|  |  |  def_object = T: EPISODIC (5.0/1.0)
|  |  |  def_object = F
|  |  |  |  absent_object = T: EPISODIC (3.0/1.0)
|  |  |  |  absent_object = F: HABITUAL (6.0/1.0)
|  |  bare_pl_subject = F: EPISODIC (784.0/94.0)
|  tense = PR
|  |  perfect = T
|  |  |  def_object = T: EPISODIC (4.0)
|  |  |  def_object = F
|  |  |  |  absent_object = T: HABITUAL (7.0/2.0)
|  |  |  |  absent_object = F
|  |  |  |  |  def_subject = T: EPISODIC (8.0/3.0)
|  |  |  |  |  def_subject = F: HABITUAL (2.0)
|  |  perfect = F: HABITUAL (126.0/18.0)
|  tense = IN
|  |  specific_temporal = T: HABITUAL (2.0)
|  |  specific_temporal = F
|  |  |  in_prep = T: HABITUAL (2.0)
|  |  |  in_prep = F: EPISODIC (45.0/8.0)
```

Figure 2. J48 Decision Tree

[2] The performance results listed for the Association Rule classifier were not evaluated using an independent validation set and hence may not be a true indicator of performance because of over-fitting of the rules to the data set.

The single largest pattern picked up the J48 classifier is for a past tense sentence without a bare plural subject which is not modified by a quantifying temporal. This combination covers 784 sentences generating a precision of 88.0% in recognizing episodic sentences. This single rule is also responsible for lowering habitual recall by 33.1%. The second most significant pattern was infinitive sentences without a temporal modifier (either quantitative or specific) and not having an '*in*'-headed prepositional phrase. This pattern covers 45 sentences generating a precision of 82.2% in recognizing episodic sentences and also lowers habitual recall by 2.8%.

For categorization of habituality, the single pattern that generated most coverage was present tense sentences which were not in perfect aspect covering a total of 126 sentences and generating a precision of 85.7%. The second most significant pattern was the presence of a quantitative temporal modifier which had a precision of 91.4% in identifying habitual sentences.

3.3 Naïve Bayes Classifier

Mitchell (1997) states that probabilistic algorithms such as Naïve Bayes are among the most effective algorithms currently known for learning to classify text documents and we have trained such a classifier in order to compare the performance with the rule-based classifiers discussed earlier. We used Weka's implementation of Naïve Bayes. As with the decision tree classifier, for smoothing of the results, we applied ten-fold cross validation while evaluating performance. In contrast to rule-based classifiers, probabilistic algorithms such as Naïve Bayes are not easily interpretable by humans (Sebastiani 2002) and hence reasoning behind the classifier performance (shown in Table 7) cannot be easily attributed. Naïve Bayes provides the highest recall of habitual sentences amongst all the classifiers we evaluated.

3.4 Individual Impact of Features

Based on the feature analysis discussed earlier, it is expected that the feature of tense is the single best discriminator amongst all the features we have considered. This is borne out of the results below where a classifier is built 1) only using tense as a discriminator 2) using all features except tense as discriminators.

	Habitual		Episodic	
Feature	**Precision**	**Recall**	**Precision**	**Recall**
Tense	82.6%	51.9%	84.4%	96.0%
Without Tense	73.4%	33.1%	79.4%	95.6%

Table 7. Impact of Tense as a feature

Performance of a J48 tree trained using all features is an improvement over just using tense as a discriminator – both the episodic and habitual precision show more than a one standard deviation improvement over the results of using just a single tense feature. Use of tense alone as a discriminator outperforms Naïve Bayes in classifying habitual sentences in terms of precision.

In order to study the impact of other features, we built classifiers excluding feature sets (i.e. excluding all subject features, excluding all temporal features etc). As expected, the largest number of incorrectly classified sentences (shown by the white bar) is when tense is excluded followed by exclusion of temporal features. This is in line with the J48 decision tree in Figure 2 where the quantitative temporal modifier feature and tense are the two most favored category discriminators. Interestingly the results show that all models are able to correctly classify roughly around the same number of episodic sentences (±6) but show greater variation in the ability to classify habitual sentences (±42). Amongst the episodic sentences, 717 sentences appear in common across all the different variations. 462 of these sentences are past tense with no aspect with a definite subject and with no temporal modifier.

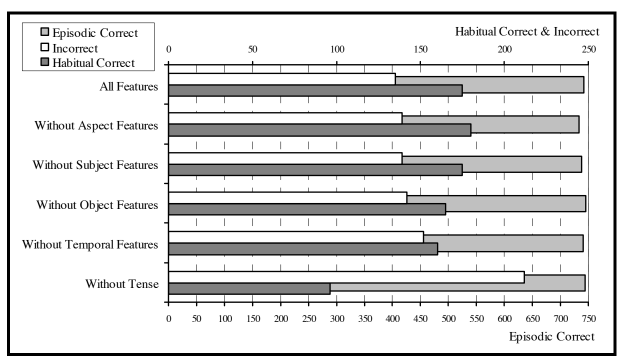

Figure 3. Impact of Feature Groups

4 Conclusion

From the results of experimentation presented earlier, we conclude that a categorization model trained using a machine learning approach against syntactic/grammatical attributes of a sentence is a viable method for discriminating between habitual and episodic sentences. The approach performs significantly better for categorizing episodic sentences as compared with habitual sentences. This is expected given that episodic sentences were more prevalent than habitual sentences in the data we evaluated – 73% of the data fell into the episodic category versus 27% in the habitual category.

Figure 4 shows the impact of varying the features used and the specific categorization algorithm on performance. Out of the different supervised approaches evaluated, use of a decision tree provided the overall best performance with an overall precision of 86.3% which was marginally higher than the overall precision of 86.1% provided by Naïve Bayes. Of all the performance measures, habitual recall was low across all classifiers with Naïve Bayes providing the highest habitual recall figure (63%).

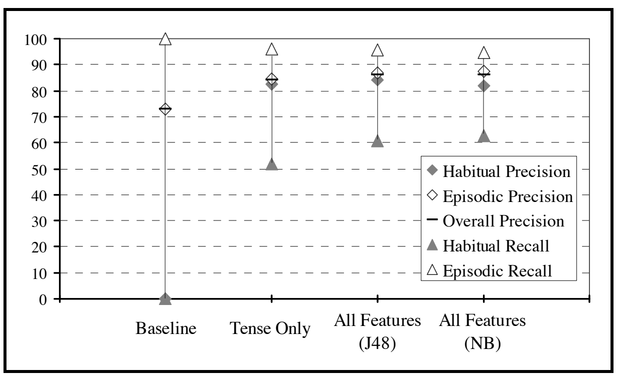

Figure 4. Comparative Performance

Among all the fourteen syntactic features which we considered, the tense of the sentence and the presence/absence of a temporal modifier stood out as having the largest impact from a supervised learning perspective. These two features (tense=present, quantificational temporal=present) were also individually the best two individual indicators for habituality. However neither of these two features covered more than 50% of the overall habitual set of sentences which to some degree explains the low habitual recall of our classifiers. With regard to aspect, the data in the Penn Treebank corpus does not present a large set of examples presenting progressive and perfect aspectual variation – and hence the impact of such features is minimized in the trained model.

It should be noted that the overall precision may not be an adequate indicator of suitability within a specific text analysis application - different problems may require varying demands of precision and recall by category. For example, an event extraction application would be better served with higher performance on the episodicity category while applications to automatically assemble knowledge base/encyclopedic repositories may require higher performance on identifying the habitual category. For applications where a supervised categorization approach is augmented by a human validator, erring on the side of higher recall numbers may be preferable for better coverage.

The robustness of this approach, if deployed within an application, is dependent on many factors such as: 1) The performance of a parser to parse and annotate raw text successfully using an annotation scheme similar to the Penn Treebank scheme, 2) The performance of the individual feature extraction processes, and 3) The nature of the classifier training data and the nature of the data that requires categorization.

4.1 Future Work

In order to improve habitual recall performance, other features other possible feature candidates that can be tied to habituality can be evaluated such as: 1) Negative modifiers (such as *never*), 2) Possible feature candidates from nested clauses, and 3) The distribution of noun phrases by definiteness/indefiniteness across a sentence.

For more comprehensive insight into category ambiguity, in addition to the unilateral category annotation scheme used in this project, it would be of interest to independently annotate the category of a sentence, its predecessor and its successor in an isolated context (i.e. disregarding wider discourse context). A weighting factor could be assigned to indicate if the annotated feels the sentence is more likely to be used in a habitual context or more likely to be used in an episodic context. This would allow the study of: 1) The impact of discourse on habituality, and 2) The flow of habitual sentences in discourse.

References

Abbott, Barbara. 2003. Definiteness and Indefiniteness. In *Handbook of Pragmatics*, eds. Laurence R. Horn and Gregory Ward, 122-149. Oxford: Blackwell.

Adrissono, Liliana, Leonardo Lesmo, Paolo Pogliano, and Paolo Terenziani. 1991. Interpretation of Definite Noun Phrases. In *Proceedings of the Twelfth International Conference on Artificial Intelligence,* eds. John Mylopoulos and Ray Reiter: 997-1002. San Mateo: Morgan Kaufmann.

Bies, Ann, Mark Ferguson, Karen Katz and Robert MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual (accessed Feb 6, 2009).

Brent, Michael R. 1991. Automatic semantic classification of verbs from their syntactic contexts: an implemented classifier for stativity. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*: 222-226.

Carlson, Gregory N. 2005. Generics, Habituals and Iteratives. In *The Encyclopedia of Language and Linguistics* 2nd edition. Elsevier.

Carlson, Gregory N. and Beverly Spejewski. 1997. Generic Passages. In *Natural Language Semantics* 5 no. 2: 101-165.

Chiercha, Gennaro. 1995. Individual-Level Predicates as Inherent Generics. *The Generic Book*, eds. Gregory N. Carlson and Francis J. Pelletier, 1-124. Chicago: University of Chicago Press.

Cowper, Elizabeth. 2003. Simplex and Complex Progressives in English and Spanish. *Workshop on Complex Predicates* – University of Toronto.

Cowper, Elizabeth. 2003. Tense, Mood and Aspect: A Feature-Geometric Approach. University of Toronto.

Dahl, Osten. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Dahl, Osten. 1995. The Marking of the Episodic/Generic Distinction Tense Aspect Systems in *The Generic Book*, eds. Gregory N. Carlson and Francis J. Pelletier, 412-426. Chicago: University of Chicago Press

Filip, Hana and Gregory N. Carlson. 1997. Sui Generis Genericity. *Penn Working Papers in Linguistics 4*, 91-110. Philadelphia: The University of Pennsylvania.

Glasbey, Sheila. 1994. Bare Plurals in Object Position: Which Verbs Fail to Give Existential Readings, and Why?. In *Indefiniteness and Plurality*, eds. Liliane Tasmowski and Svetlana Vogeleer. Amsterdam: Benjamins.

Kearns, Katherine S. 1991. The Semantics of the English Progressive. PhD. diss., MIT.

Krifka, Manfred, Francis J. Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link and Gennaro Chierchia. 1995. Genericity: An Introduction. *The Generic Book*, eds. Gregory N. Carlson and Francis J. Pelletier, 1-124. Chicago: University of Chicago Press.

Mathew, Thomas. 2009. Supervised Categorization of Habitual versus Episodic Sentences. Master's thesis, Georgetown University.

Mitchell, Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert McIntyre, Ann Bies, Mark Ferguson, Karen Katz and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *ARPA Human Language Technology Workshop*: 114-119. Morristown: Association for Computational Linguistics.

Mitchell, Tom M. 1997. *Machine Learning*. Singapore: McGraw-Hill.

Rimell, Laura. 2004. Habitual Sentences and Generic Quantification. In *Proceedings of the 23rd West Coast Conference on Formal Linguistics,* eds. Vineeta Chand, Ann Kelleher, Angelo J. Rodríguez, and Benjamin Schmeiser: 663-676. Somerville: Cascadilla Press.

Rohde, Douglas L. T. 2000. Extracting Syntax Statistics from Large Corpora of Written English. Carnegie Mellon University.

Scheffer, Tobias. 1994. Finding Association Rules That Trade Support Optimally against Confidence. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*: 424-435. Freiburg: Springer-Verlag.

Sebastiani, Fabrizio. 2002. Machine Learning in Automated Text Categorization. In *ACM Computing Surveys* 34, no. 1: 1-47. New York: ACM.

Siegel, Eric V. 1999. Corpus-Based Linguistic Indicators for Aspectual Classification. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*: 112-119. San Francisco: Morgan Kaufmann.

Suh, S. 2006. Extracting Generic Statements for the Semantic Web. Master's thesis, The University of Edinburgh.

Tagliamonte, Sali and Lawrence, Helen. 2000. "I used to dance, but I don't dance now.": The Habitual Past in English. In *Journal of English Linguistics* 28, no. 4: 323-353.