PREDICTING INTONATIONAL PHRASING FROM TEXT

Michelle Q. Wang
Churchill College
Cambridge University
Cambridge UK

Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Abstract

Determining the relationship between the intonational characteristics of an utterance and other features inferable from its text is important both for speech recognition and for speech synthesis. This work investigates the use of text analysis in predicting the location of intonational phrase boundaries in natural speech, through analyzing 298 utterances from the DARPA Air Travel Information Service database. For statistical modeling, we employ Classification and Regression Tree (CART) techniques. We achieve success rates of just over 90%, representing a major improvement over other attempts at boundary prediction from unrestricted text.[1]

Introduction

The relationship between the intonational phrasing of an utterance and other features which can be inferred from its transcription represents an important source of information for speech synthesis and speech recognition. In synthesis, more natural intonational phrasing can be assigned if text analysis can predict human phrasing performance. In recognition, better calculation of probable word durations is possible if the phrase-final-lengthening that precedes boundary sites can be predicted. Furthermore, the association of intonational features with syntactic and acoustic information can also be used to reduce the number of sentence hypotheses under consideration.

Previous research on the location of intonational boundaries has largely focussed on the relationship between these prosodic boundaries and syntactic constituent boundaries. While current research acknowledges the role that semantic and discourse-level information play in boundary assignment, most authors assume that syntactic configuration provides the basis for prosodic 'defaults' that may be overridden by semantic or discourse considerations. While most interest in boundary prediction has been focussed on synthesis (Gee and Grosjean, 1983; Bachenko and Fitzpatrick, 1990), currently there is considerable interest in predicting boundaries to aid recognition (Ostendorf et al., 1990; Steedman, 1990). The most successful empirical studies in boundary location have investigated how phrasing can disambiguate potentially syntactically ambiguous utterances in read speech (Lehiste, 1973; Ostendorf et al., 1990). Analysis based on corpora of natural speech (Altenberg, 1987) have so far reported very limited success and have assumed the availability of syntactic, semantic, and discourse-level information well beyond the capabilities of current NL systems to provide.

To address the question of how boundaries are assigned in natural speech — as well as the need for classifying boundaries from information that can be extracted automatically from text — we examined a multi-speaker corpus of spontaneous elicited speech. We wanted to compare performance in the prediction of intonational boundaries from information available through simple techniques of text analysis, to performance using information currently available only come from hand labeling of transcriptions. To this end, we selected potential boundary predictors based upon hypotheses derived from our own observations and from previous theoretical and practical studies of boundary location. Our corpus for this investigation is 298 sentences from approximately 770 sentences of the Texas Instruments-collected portion of the DARPA Air Travel Information Service (ATIS) database(DAR, 1990). For statistical modeling, we employ classification and regression tree techniques (CART) (Brieman et al., 1984), which provide cross-validated decision trees for boundary classification. We obtain (cross-validated) success rates of 90% for both automatically-generated information and hand-

[1] We thank Michael Riley for helpful discussions. Code implementing the CART techniques employed here was written by Michael Riley and Daryl Pregibon. Part-of-speech tagging employed Ken Church's tagger, and syntactic analysis used Don Hindle's parser, Fidditch.

labeled data on this sample, which represents a major improvement over previous attempts to predict intonational boundaries for spontaneous speech and equals or betters previous (hand-crafted) algorithms tested for read speech.

Intonational Phrasing

Intuitively, intonational phrasing divides an utterance into meaningful 'chunks' of information (Bolinger, 1989). Variation in phrasing can change the meaning hearers assign to tokens of a given sentence. For example, interpretation of a sentence like '*Bill doesn't drink because he's unhappy.*' will change, depending upon whether it is uttered as one phrase or two. Uttered as a single phrase, this sentence is commonly interpreted as conveying that Bill does indeed drink — but the cause of his drinking is not his unhappiness. Uttered as two phrases, it is more likely to convey that Bill does *not* drink — and the reason for his abstinence is his unhappiness.

To characterize this phenomenon phonologically, we adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert, 1980). In this view, two levels of phrasing are significant in English intonational structure. Both types are composed of sequences of high and low tones in the FUNDAMENTAL FREQUENCY (f0) contour. An INTERMEDIATE (or minor) PHRASE consists of one or more PITCH ACCENTS (local f0 minima or maxima) plus a PHRASE ACCENT (a simple high or low tone which controls the pitch from the last pitch accent of one intermediate phrase to the beginning of the next intermediate phrase or the end of the utterance). INTONATIONAL (or major) PHRASES consist of one or more intermediate phrases plus a final BOUNDARY TONE, which may also be high or low, and which occurs at the end of the phrase. Thus, an intonational phrase boundary necessarily coincides with an intermediate phrase boundary, but not vice versa.

While phrase boundaries are perceptual categories, they are generally associated with certain physical characteristics of the speech signal. In addition to the tonal features described above, phrases may be identified by one of more of the following features: pauses (which may be filled or not), changes in amplitude, and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization of that syllable and perhaps preceding syllables). In general, major phrase boundaries tend to be associated with longer pauses, greater tonal changes, and more final lengthening than minor boundaries.

The Experiments

The Corpus and Features Used in Analysis

The corpus used in this analysis consists of 298 utterances (24 minutes of speech from 26 speakers) from the speech data collected by Texas Instruments for the DARPA Air Travel Information System (ATIS) spoken language system evaluation task. In a Wizard-of-Oz simulation, subjects were asked to make travel plans for an assigned task, providing spoken input and receiving teletype output. The quality of the ATIS corpus is extremely diverse. Speaker performance ranges from close to isolated-word speech to exceptional fluency. Many utterances contain hesitations and other disfluencies, as well as long pauses (greater than 3 sec. in some cases).

To prepare this data for analysis, we labeled the speech prosodically by hand, noting location and type of intonational boundaries and presence or absence of pitch accents. Labeling was done from both the waveform and pitchtracks of each utterance. Each label file was checked by several labelers. Two levels of boundary were labeled; in the analysis presented below, however, these are collapsed to a single category.

We define our data points to consist of all potential boundary locations in an utterance, defined as each pair of adjacent words in the utterance $< w_i, w_j >$, where w_i represents the word to the left of the potential boundary site and w_j represents the word to the right.[2] Given the variability in performance we observed among speakers, an obvious variable to include in our analysis is speaker identity. While for applications to speaker-independent recognition this variable would be uninstantiable, we nonetheless need to determine how important speaker idiosyncracy may be in boundary location. We found no significant increase in predictive power when this variable is used. Thus, results presented below are speaker-independent.

One easily obtainable class of variable involves temporal information. Temporal variables include utterance and phrase duration, and distance of the

[2] See the appendix for a partial list of variables employed, which provides a key to the node labels for the prediction trees presented in Figures 1 and 2.

potential boundary from various strategic points in the utterance. Although it is tempting to assume that phrase boundaries represent a purely intonational phenomenon, it is possible that processing constraints help govern their occurrence. That is, longer utterances may tend to include more boundaries. Accordingly, we measure the length of each utterance both in seconds and in words. The distance of the boundary site from the beginning and end of the utterance is another variable which appears likely to be correlated with boundary location. The tendency to end a phrase may also be affected by the position of the potential boundary site in the utterance. For example, it seems likely that positions very close to the beginning or end of an utterance might be unlikely positions for intonational boundaries. We measure this variable too, both in seconds and in words. The importance of phrase length has also been proposed (Gee and Grosjean, 1983; Bachenko and Fitzpatrick, 1990) as a determiner of boundary location. Simply put, it seems may be that consecutive phrases have roughly equal length. To capture this, we calculate the elapsed distance from the last boundary to the potential boundary site, divided by the length of the last phrase encountered, both in time and words. To obtain this information automatically would require us to factor prior boundary predictions into subsequent predictions. While this would be feasible, it is not straightforward in our current classification strategy. So, to test the utility of this information, we have used observed boundary locations in our current analysis.

As noted above, syntactic constituency information is generally considered a good predictor of phrasing information (Gee and Grosjean, 1983; Selkirk, 1984; Marcus and Hindle, 1985; Steedman, 1990). Intuitively, we want to test the notion that some constituents may be more or less likely than others to be internally separated by intonational boundaries, and that some syntactic constituent boundaries may be more or less likely to coincide with intonational boundaries. To test the former, we examine the class of the lowest node in the parse tree to dominate both w_i and w_j, using Hindle's parser, Fidditch (1989) To test the latter we determine the class of the highest node in the parse tree to dominate w_i, but not w_j, and the class of the highest node in the tree to dominate w_j but not w_i. Word class has also been used often to predict boundary location, particularly in text-to-speech. The belief that phrase boundaries rarely occur after function words forms the basis for most algorithms used to assign intonational phrasing for text-to-speech. Furthermore, we might expect that some words, such as prepositions and determiners, for example, do not constitute the typical end to an intonational phrase. We test these possibilities by examining part-of-speech in a window of four words surrounding each potential phrase break, using Church's part-of-speech tagger (1988).

Recall that each intermediate phrase is composed of one or more pitch accents plus a phrase accent, and each intonational phrase is composed of one or more intermediate phrases plus a boundary tone. Informal observation suggests that phrase boundaries are more likely to occur in some accent contexts than in others. For example, phrase boundaries between words that are deaccented seem to occur much less frequently than boundaries between two accented words. To test this, we look at the pitch accent values of w_i and w_j for each $< w_i, w_j >$, comparing observed values with predicted pitch accent information obtained from (Hirschberg, 1990).

In the analyses described below, we employ varying combinations of these variables to predict intonational boundaries. We use classification and regression tree techniques to generate decision trees automatically from variable values provided.

Classification and Regression Tree Techniques

Classification and regression tree (CART) analysis (Brieman et al., 1984) generates decision trees from sets of continuous and discrete variables by using set of splitting rules, stopping rules, and prediction rules. These rules affect the internal nodes, subtree height, and terminal nodes, respectively. At each internal node, CART determines which factor should govern the forking of two paths from that node. Furthermore, CART must decide which values of the factor to associate with each path. Ideally, the splitting rules should choose the factor and value split which minimizes the prediction error rate. The splitting rules in the implementation employed for this study (Riley, 1989) approximate optimality by choosing at each node the split which minimizes the prediction error rate on the training data. In this implementation, all these decisions are binary, based upon consideration of each possible binary partition of values of categorical variables and consideration of different cut-points for values of continuous variables.

Stopping rules terminate the splitting process at each internal node. To determine the best tree, this implementation uses two sets of stopping rules. The first set is extremely conservative, resulting in an overly large tree, which usually lacks the generality necessary to account for data outside of the training set. To compensate, the second rule set forms a sequence of subtrees. Each tree is grown on a sizable fraction of the training data and tested on the remaining portion. This step is repeated until the tree has been grown and tested on all of the data. The stopping rules thus have access to cross-validated error rates for each subtree. The subtree with the lowest rates then defines the stopping points for each path in the full tree. Trees described below all represent cross-validated data.

The prediction rules work in a straightforward manner to add the necessary labels to the terminal nodes. For continuous variables, the rules calculate the mean of the data points classified together at that node. For categorical variables, the rules choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. In this implementation, the deviation for continuous variables is the sum of the squared error for the observations. The deviation for categorical variables is simply the number of misclassified observations.

Results

In analyzing boundary locations in our data, we have two goals in mind. First, we want to discover the extent to which boundaries can be predicted, given information which can be generated automatically from the text of an utterance. Second, we want to learn how much predictive power can be gained by including additional sources of information which, at least currently, cannot be generated automatically from text. In discussing our results below, we compare predictions based upon automatically inferable information with those based upon hand-labeled data.

We employ four different sets of variables during the analysis. The first set includes observed phonological information about pitch accent and prior boundary location, as well as automatically obtainable information. The success rate of boundary prediction from the variable set is extremely high, with correct cross-validated classification of 3330 out of 3677 potential boundary sites — an overall success rate of 90% (Figure 1). Furthermore, there are only five decision points in the tree. Thus, the tree represents a clean, simple model of phrase boundary prediction, assuming accurate phonological information.

Turning to the tree itself, we that the ratio of current phrase length to prior phrase length is very important in boundary location. This variable alone (assuming that the boundary site occurs before the end of the utterance) permits correct classification of 2403 out of 2556 potential boundary sites. Occurrence of a phrase boundary thus appears extremely unlikely in cases where its presence would result in a phrase less than half the length of the preceding phrase. The first and last decision points in the tree are the most trivial. The first split indicates that utterances virtually always end with a boundary — rather unsurprising news. The last split shows the importance of distance from the beginning of the utterance in boundary location; boundaries are more likely to occur when more than 2 $\frac{1}{2}$ seconds have elapsed from the start of the utterance.[3] The third node in the tree indicates that noun phrases form a tightly bound intonational unit. The fourth split in 1 shows the role of accent context in determining phrase boundary location. If w_i is not accented, then it is unlikely that a phrase boundary will occur after it.

The significance of accenting in the phrase boundary classification tree leads to the question of whether or not predicted accents will have a similar impact on the paths of the tree. In the second analysis, we substituted predicted accent values for observed values. Interestingly, the success rate of the classification remained approximately the same, at 90%. However, the number of splits in the resultant tree increased to nine and failed to include the accenting of w_i as a factor in the classification. A closer look at the accent predictions themselves reveals that the majority of misclassifications come from function words preceding a boundary. Although the accent prediction algorithm predicted that these words would be deaccented, they were in fact accented. This appears to be an idiosyncracy of the corpus; such words generally occurred before relatively long pauses. Nevertheless, classification succeeds well in the absence of accent information, perhaps suggesting that accent values may themselves be highly correlated with other variables. For example, both pitch accent and boundary location appear sensitive to location of prior intonational boundaries and part-of-speech.

[3] This fact may be idiosyncratic to our data, given the fact that we observed a trend towards initial hesitations.

In the third analysis, we eliminate the dynamic boundary percentage measure. The result remains nearly as good as before, with a success rate of 89%. The proposed decision tree confirms the usefulness of observed accent status of w_i in boundary prediction. By itself (again assuming that the potential boundary site occurs before the end of the utterance), this factor accounts for 1590 out of 1638 potential boundary site classifications. This analysis also confirms the strength of the intonational ties among the components of noun phrases. In this tree, 536 out of 606 potential boundary sites receive final classification from this feature.

We conclude our analysis by producing a classification tree that uses automatically-inferrable information alone. For this analysis we use predicted accent values instead of observed values and omit boundary distance percentage measures. Using binary-valued accented predictions (i.e., are $< w_i, w_j >$ accented or not), we obtain a success rate for boundary prediction of 89%, and using a four-valued distinction for predicted accented (cliticized, deaccented, accented, 'NA') we increased this to 90%. The tree in Figure 2) presents the latter analysis.

Figure 2 contains more nodes than the trees discussed above; more variables are used to obtain a similar classification percentage. Note that accent predictions are used trivially, to indicate sentence-final boundaries (*ra*='NA'). In figure 1, this function was performed by distance of potential boundary site from end of utterance (*et*). The second split in the new tree does rely upon temporal distance — this time, distance of boundary site from the beginning of the utterance. Together these measurements correctly predict nearly forty percent of the data (38.2%). Th classifier next uses a variable which has not appeared in earlier classifications — the part-of-speech of w_j. In 2, in the majority of cases (88%) where w_j is a function word other than 'to,' 'in,' or a conjunction (true for about half of potential boundary sites), a boundary does not occur. Part-of-speech of w_i and type of constituent dominating w_i but not w_j are further used to classify these items. This portion of the classification is reminiscent of the notion of 'function word group' used commonly in assigning prosody in text-to-speech, in which phrases are defined, roughly, from one function word to the next. Overall rate of the utterance and type of utterance appear in the tree, in addition to part-of-speech and constituency information, and distance of potential boundary site from beginning and end of utterance. In general, results of this first stage of analysis suggest — encouragingly — that there is considerable redundancy in the features predicting boundary location: when some features are unavailable, others can be used with similar rates of success.

Discussion

The application of CART techniques to the problem of predicting and detecting phrasing boundaries not only provides a classification procedure for predicting intonational boundaries from text, but it increases our understanding of the importance of several among the numerous variables which might plausibly be related to boundary location. In future, we plan to extend the set of variables for analysis to include counts of stressed syllables, automatic NP-detection (Church, 1988), MUTUAL INFORMATION, GENERALIZED MUTUAL INFORMATION scores can serve as indicators of intonational phrase boundaries (Magerman and Marcus, 1990).

We will also examine possible interactions among the statistically important variables which have emerged from our initial study. CART techniques have worked extremely well at classifying phrase boundaries and indicating which of a set of potential variables appear most important. However, CART's step-wise treatment of variables, optimization heuristics, and dependency on binary splits obscure the possible relationships that exist among the various factors. Now that we have discovered a set of variables which do well at predicting intonational boundary location, we need to understand just how these variables interact.

References

Bengt Altenberg. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English*. Lund University Press, Lund.

J. Bachenko and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*. To appear.

Dwight Bolinger. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse*. Edward Arnold, London.

Leo Brieman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Wadsworth & Brooks, Monterrey CA.

K. W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin. Association for Computational Linguistics.

DARPA. 1990. *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley PA, June.

J. P. Gee and F. Grosjean. 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458.

D. M. Hindle. 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting*, pages 118–125, Vancouver. Association for Computational Linguistics.

Julia Hirschberg. 1990. Assigning pitch accent in synthetic speech: The given/new distinction and deaccentability. In *Proceedings of the Seventh National Conference*, pages 952–957, Boston. American Association for Artificial Intelligence.

I. Lehiste. 1973. Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7:197–222.

David M. Magerman and Mitchel P. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90*, pages 984–989. American Association for Artifical Intelligence.

Mitchell P. Marcus and Donald Hindle. 1985. A computational account of extra categorial elements in japanese. In *Papers presented at the First SDF Workshop in Japanese Syntax.* System Development Foundation.

M. Ostendorf, P. Price, J. Bear, and C. W. Wightman. 1990. The use of relative duration in syntactic disambiguation. In *Proceedings of the DARPA Speech and Natural Language Workshop.* Morgan Kaufmann, June.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology, September.

Michael D. Riley. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings.* DARPA Speech and Natural Language Workshop, October.

E. Selkirk. 1984. *Phonology and Syntax.* MIT Press, Cambridge MA.

M. Steedman. 1990. Structure and intonation in spoken language understanding. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics.*

Appendix: Key to Figures

for each potential boundary, $< w_i, w_j >$	
type	utterance type
tt	total # seconds in utterance
tw	total # words in utterance
st	distance (sec.) from start to w_j
et	distance (sec.) from w_j to end
sw	distance (words) from start to w_j
ew	distance (words) from w_j to end
la	is w_i accented or not/ or, cliticized, deaccented, accented
ra	is w_j accented or not/ or, cliticized, deaccented, accented
per	[distance (words) from last boundary]/ [length (words) of last phrase]
tper	[distance (sec.) from last boundary]/ [length (sec.) of last phrase]
j{1-4}	part-of-speech of $w_{i-1,i,j,j+1}$ v = verb b = be-verb m = modifier f = fn word n = noun p = preposition w = WH
f{slr}	category of s = smallest constit dominating w_i,w_j l = largest constit dominating w_i, not w_j r = largest constit dominating w_j, not w_i m = modifier d = determiner v = verb p = preposition w = WH n = noun s = sentence f = fn word

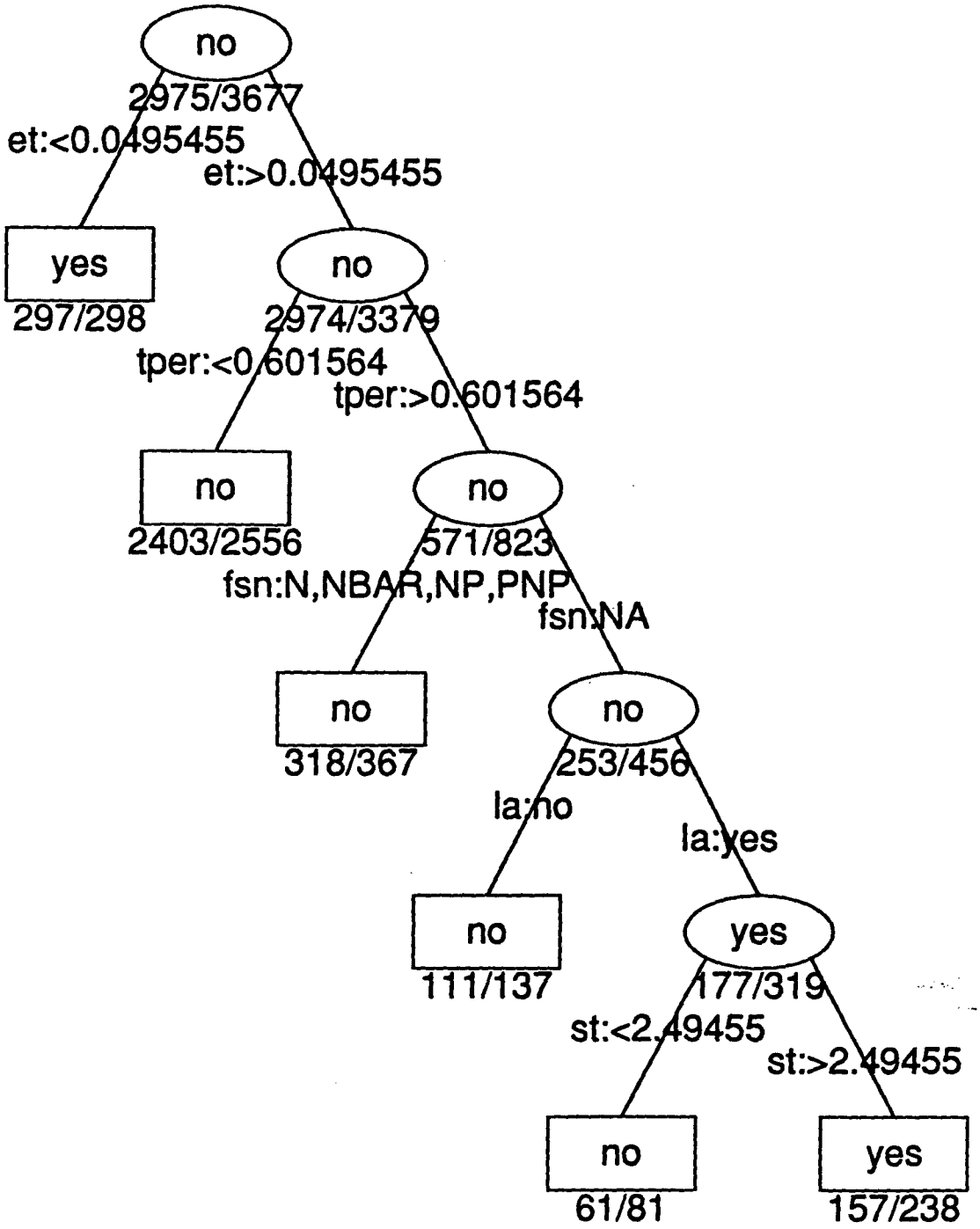

Figure 1: Predictions from Automatically-Acquired and Observed Data, 90%

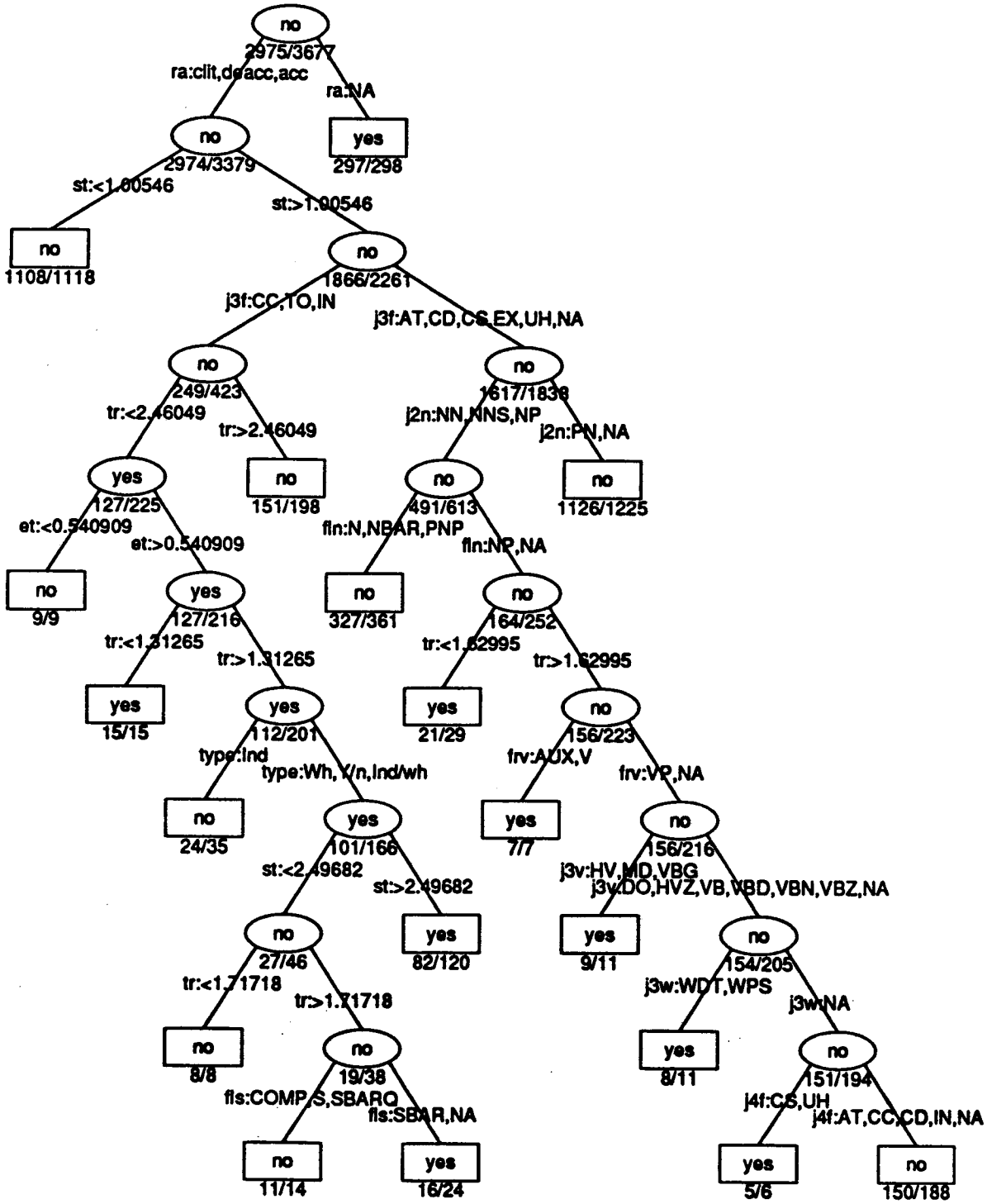

Figure 2: Phrase Boundary Predictions from Automatically-Inferred Information, 90%